1. Chapter 3 - Probability Topics
    1. Introduction -- Probability Topics -- MtRoyal - Version2016RevA
    2. Terminology -- Probability Topics -- MtRoyal - Version2016RevA
    3. Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA
    4. Two Basic Rules of Probability
    5. Contingency Tables and Tree Diagrams -- Probability Topics -- MtRoyal - Version2016RevA
2. Chapter 4 - Binomial Distribution
    1. Introduction to Discrete Probability Distributions - MRU - C Lemieux (2017)
    2. Binomial distribution - MRU - C Lemieux
3. Chapter 5 - Normal Distribution
    1. Introduction -- The Normal Distribution -- Mt Royal University -- Version 2016RevA
    2. The Standard Normal Distribution-- The Normal Distribution --MRU - C Lemieux
    3. Using the Normal Distribution-- The Normal Distribution -- MRU - C Lemieux
4. Chapter 6 - Sampling Distribution
    1. Introduction - Sampling distributions - MRU - C Lemieux
    2. Introduction to Sampling Distributions
    3. Constructing empirical sampling distributions - MRU - C Lemieux
    4. Central Limit Theorem - MRU - C Lemieux
    5. A series of examples - Calculating probabilities for sampling distributions - MRU - C Lemieux

Introduction -- Probability Topics -- MtRoyal - Version2016RevA class="introduction"

Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

Note:
Chapter objective
By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the addition and multiplication rules.

- Construct and interpret contingency tables and tree diagrams.
- Understand the difference between likely and unlikely events.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Terminology -- Probability Topics -- MtRoyal - Version2016RevA

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter *S* is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like *A* and *B* represent events. For example, if the experiment is to flip one fair coin, event *A* might be getting at most one head. The probability of an event *A* is written $P(A)$.

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between zero and one, inclusive** (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event *A* can never happen. $P(A) = 1$ means the event *A* always happens. $P(A) = 0.5$ means that event *A* has a 50% chance of happening. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event *A*

and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is {*HH*, *TH*, *HT*, *TT*} where *T* = tails and *H* = heads. The sample space has four outcomes. *A* = getting one head. There are two outcomes that meet this condition {*HT*, *TH*}, so $P(A) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event *E* = rolling a number that is at least five. There are two outcomes {5, 6}. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

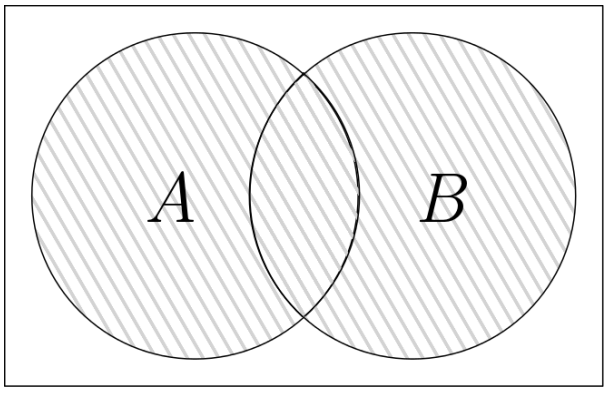
This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces.

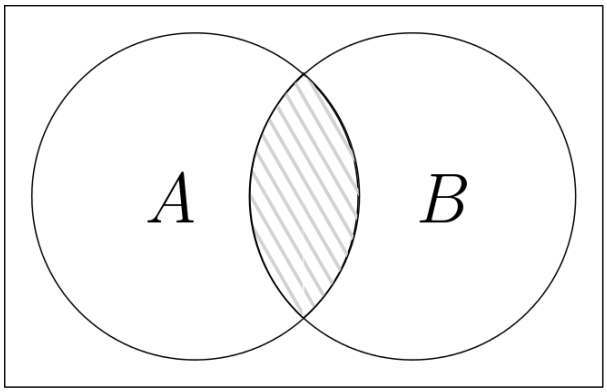
Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

A key concept in probability is whether an event is **likely** or **unlikely**. A likely event is an event that has a good chance of happening, while an unlikely event is rare. For example, it is likely to snow in Calgary in the winter, but it is unlikely to snow in Calgary in the summer (it can happen, but it would be a rare or strange event). In general, in statistics, unlikely events usually have a probability of less than 1% of happening. Likely events usually have a probability of greater than 10% of happening. If the probability of the event is between 1% and 10%, it is up to the statistician or researcher to make a call to determine whether it is likely or unlikely.

"OR" Event:
An outcome is in the event *A* OR *B* if the outcome is in *A* or is in *B* or is in both *A* and *B*. For example, let *A* = {1, 2, 3, 4, 5} and *B* = {4, 5, 6, 7, 8}. *A* OR *B* = {1, 2, 3, 4, 5, 6, 7, 8}. Notice that 4 and 5 are NOT listed twice.



"AND" Event:
An outcome is in the event *A* AND *B* if the outcome is in both *A* and *B* at the same time. For example, let *A* and *B* be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then *A* AND *B* = {4, 5}.



The **complement** of event *A* is denoted *A′* (read "*A* prime"). *A′* consists of all outcomes that are **NOT** in *A*. Notice that *P*(*A*) + *P*(*A′*) = 1. For example, let *S* = {1, 2, 3, 4, 5, 6} and let *A* = {1, 2, 3, 4}. Then, *A′* = {5, 6}. *P*(*A*) = $\frac{4}{6}$, *P*(*A′*) = $\frac{2}{6}$, and *P*(*A*) + *P*(*A′*) = $\frac{4}{6} + \frac{2}{6} = 1$

The **conditional probability** of *A* given *B* is written *P*(*A*|*B*). *P*(*A*|*B*) is the probability that event *A* will occur given that the event *B* has already occurred. **A conditional reduces the sample space**. We calculate the probability of *A* from the reduced sample space *B*. The formula to calculate *P*(*A*|*B*) is *P*(*A*|*B*) = $\frac{P(A \text{ AND } B)}{P(B)}$ where *P*(*B*) is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space *S* = {1, 2, 3, 4, 5, 6}. Let *A* = face is 2 or 3 and *B* = face is even (2, 4, 6). To calculate *P*(*A*|*B*), we count the number of outcomes 2 or 3 in the sample space *B* = {2, 4, 6}. Then we divide that by the number of outcomes *B* (rather than *S*).

We get the same result by using the formula. Remember that *S* has six outcomes.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{(\text{the number of outcomes that are 2 or 3 and even in } S)}{6}}{\frac{(\text{the number of outcomes that are even in } S)}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Odds

The odds of an event presents the probability as a ratio of success to failure. This is common in various gambling formats. Mathematically, the odds of an event can be defined as:

Equation:

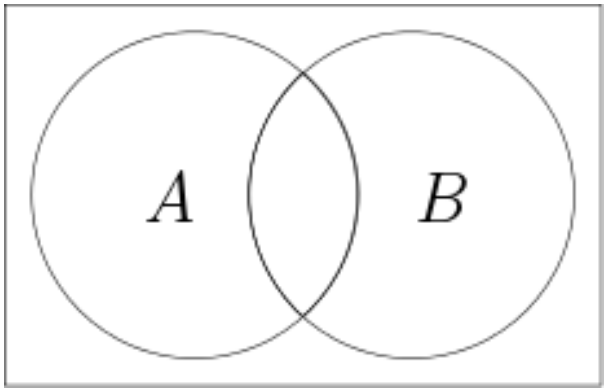
$$\frac{P(A)}{1 - P(A)}$$

where $P(A)$ is the probability of success and of course $1 - P(A)$ is the probability of failure. Odds are always quoted as "numerator to denominator," e.g. 2 to 1. Here the probability of winning is twice that of losing; thus, the probability of winning is 0.66. A probability of winning of 0.60 would generate odds in favor of winning of 3 to 2. While the calculation of odds can be useful in gambling venues in determining payoff amounts, it is not helpful for understanding probability or statistical theory.

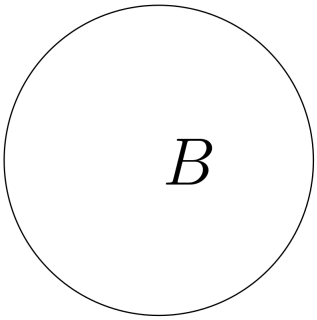
Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

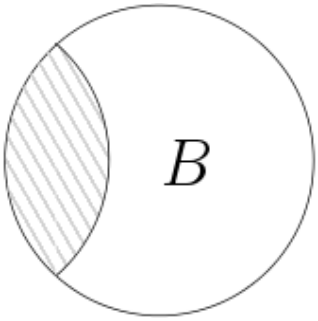
If the sample space is



then $P(A|B)$ is found by looking only at events that involved B:



and within *B* looking at the portion that involve *A*:



That portion is clearly the intersection of *A* and *B*.

Example:
Exercise:

Problem:

The sample space *S* is the whole numbers starting at one and less than 20.

a. *S* = _____

Let event *A* = the even numbers and event *B* = numbers greater than 13.

b. *A* = _____, *B* = _____

c. *P*(*A*) = _____, *P*(*B*) = _____

d. A AND B = _____, A OR B = _____
e. $P(A \text{ AND } B)$ = _____, $P(A \text{ OR } B)$ = _____
f. A' = _____, $P(A')$ = _____
g. $P(A) + P(A')$ = _____
h. $P(A|B)$ = _____, $P(B|A)$ = _____; are the probabilities equal?

Solution:

a. S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19}
b. A = {2, 4, 6, 8, 10, 12, 14, 16, 18}, B = {14, 15, 16, 17, 18, 19}
c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$
d. A AND B = {14,16,18}, A OR B = {2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19}
e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$
f. A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19; $P(A') = \frac{10}{19}$
g. $P(A) + P(A') = 1$ ($\frac{9}{19} + \frac{10}{19} = 1$)
h. $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{ AND } B)}{P(A)} = \frac{3}{9}$, No

Note:
Try It
Exercise:

Problem:

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. S = _____

Let event A = the sum is even and event B = the first number is prime.

b. A = _____, B = _____
c. $P(A)$ = _____, $P(B)$ = _____
d. A AND B = _____, A OR B = _____
e. $P(A \text{ AND } B)$ = _____, $P(A \text{ OR } B)$ = _____
f. B' = _____, $P(B')$ = _____
g. $P(A) + P(A')$ = _____
h. $P(A|B)$ = _____, $P(B|A)$ = _____; are the probabilities equal?

Solution:

a. S = {(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}
b. A = {(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)}

   B = {(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}
c. $P(A) = \frac{1}{2}$, $P(B) = \frac{2}{3}$
d. A AND B = {(2,2), (2,4), (3,1), (3,3)}

   A OR B = {(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}
e. $P(A \text{ AND } B) = \frac{1}{3}$, $P(A \text{ OR } B) = \frac{5}{6}$
f. B' = {(1,1), (1,2), (1,3), (1,4)}, $P(B') = \frac{1}{3}$
g. $P(B) + P(B') = 1$
h. $P(A|B) = \frac{P(A \text{AND} B)}{P(B)} = \frac{1}{2}$, $P(B|A) = \frac{P(A \text{AND} B)}{P(B)} = \frac{2}{3}$, No.

Example:
Exercise:

Problem:

A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

a. Event T = the outcome is two.
b. Event A = the outcome is an even number.
c. Event B = the outcome is less than four.
d. The complement of A.
e. A GIVEN B
f. B GIVEN A
g. A AND B
h. A OR B
i. A OR B'
j. Event N = the outcome is a prime number.
k. Event I = the outcome is seven.

Solution:

a. $T = \{2\}$, $P(T) = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$
c. $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$
d. $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$
e. $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$
f. $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$
g. A AND $B = \{2\}$, $P(A \text{ AND } B) = \frac{1}{6}$
h. A OR $B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = \frac{5}{6}$
i. A OR $B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = \frac{2}{3}$
j. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7) = 0$.

Example:
[link] describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	**Left-handed**
Males	43	9
Females	44	4

Exercise:

Problem:

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

a. $P(M)$
b. $P(F)$
c. $P(R)$
d. $P(L)$
e. $P(M \text{ AND } R)$
f. $P(F \text{ AND } L)$
g. $P(M \text{ OR } F)$
h. $P(M \text{ OR } R)$
i. $P(F \text{ OR } L)$
j. $P(M')$
k. $P(R|M)$
l. $P(F|L)$
m. $P(L|F)$

Solution:

a. $P(M) = 0.52$
b. $P(F) = 0.48$
c. $P(R) = 0.87$
d. $P(L) = 0.13$
e. $P(M \text{ AND } R) = 0.43$
f. $P(F \text{ AND } L) = 0.04$
g. $P(M \text{ OR } F) = 1$
h. $P(M \text{ OR } R) = 0.96$
i. $P(F \text{ OR } L) = 0.57$
j. $P(M') = 0.48$
k. $P(R|M) = 0.8269$ (rounded to four decimal places)
l. $P(F|L) = 0.3077$ (rounded to four decimal places)
m. $P(L|F) = 0.0833$

References

"Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Chapter Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Exercise:

Problem:

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Solution:

a. $P(L') = P(S)$
b. $P(M \text{ OR } S)$
c. $P(F \text{ AND } L)$
d. $P(M|L)$
e. $P(L|M)$
f. $P(S|F)$
g. $P(F|L)$
h. $P(F \text{ OR } L)$
i. $P(M \text{ AND } S)$
j. $P(F)$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.
Let H = the event of getting a hat.
Let N = the event of getting a noisemaker.
Let F = the event of getting a finger trap.
Let C = the event of getting a bag of confetti.

Exercise:

Problem: Find $P(H)$.

Exercise:

Problem: Find $P(N)$.

Solution:

$P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise:

Problem: Find $P(F)$.

Exercise:

Problem: Find $P(C)$.

Solution:

$P(C) = \frac{5}{42} = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.
Let B = the event of getting a blue jelly bean
Let G = the event of getting a green jelly bean.
Let O = the event of getting an orange jelly bean.
Let P = the event of getting a purple jelly bean.
Let R = the event of getting a red jelly bean.
Let Y = the event of getting a yellow jelly bean.

Exercise:

Problem: Find $P(B)$.

Exercise:

Problem: Find $P(G)$.

Solution:

$P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise:

Problem: Find $P(P)$.

Exercise:

Problem: Find $P(R)$.

Solution:

$P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise:

Problem: Find $P(Y)$.

Exercise:

Problem: Find $P(O)$.

Solution:

$P(O) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).
Let A = the event that a country is in Asia.
Let E = the event that a country is in Europe.
Let F = the event that a country is in Africa.
Let N = the event that a country is in North America.
Let O = the event that a country is in Oceania.
Let S = the event that a country is in South America.

Exercise:

Problem: Find $P(A)$.

Exercise:

Problem: Find $P(E)$.

Solution:

$P(E) = \frac{47}{194} = 0.24$

Exercise:

Problem: Find $P(F)$.

Exercise:

Problem: Find $P(N)$.

Solution:

$P(N) = \frac{23}{194} = 0.12$

Exercise:

Problem: Find $P(O)$.

Exercise:

Problem: Find $P(S)$.

Solution:

$P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise:

Problem:

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise:

Problem:

What is the probability of drawing a club in a standard deck of 52 cards?

Solution:

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercise:

Problem:

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise:

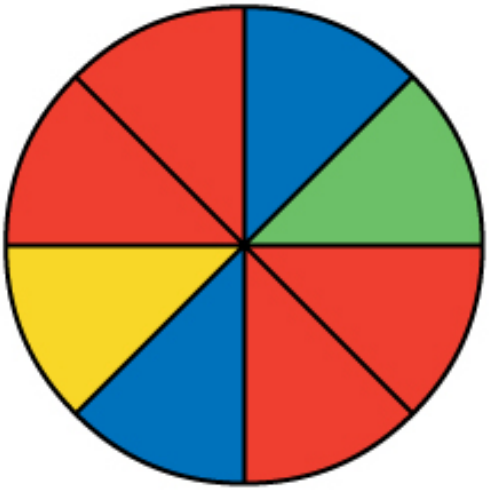
Problem:

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Solution:

$\frac{3}{6} = \frac{1}{2} = 0.5$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.



Let *B* = the event of landing on blue.
Let *R* = the event of landing on red.
Let *G* = the event of landing on green.
Let *Y* = the event of landing on yellow.

Exercise:

Problem: If you land on *Y*, you get the biggest prize. Find *P*(*Y*).

Exercise:

Problem: If you land on red, you don't get a prize. What is *P*(*R*)?

Solution:

$P(R) = \frac{4}{8} = 0.5$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.
Let *I* = the event that a player in an infielder.
Let *O* = the event that a player is an outfielder.
Let *H* = the event that a player is a great hitter.
Let *N* = the event that a player is not a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is not an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Solution:

$P(O \text{ OR } H)$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Solution:

$P(H|I)$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise:

Problem:

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Solution:

$P(N|O)$

Exercise:

Problem:

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Solution:

$P(I \text{ OR } N)$

Exercise:

Problem:

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder.

Solution:

$P(I)$

Exercise:

Problem: What is the word for the set of all possible outcomes?

Exercise:

Problem: What is conditional probability?

Solution:

The likelihood that an event will occur given that another event has already occurred.

Exercise:

Problem:

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book
Let F = event that book is fiction
Let N = event that book is nonfiction
What is the sample space?

Exercise:

Problem:

What is the sum of the probabilities of an event and its complement?

Solution:

1

Use the following information to answer the next two exercises. You are

rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

Exercise:

Problem: What does $P(E|M)$ mean in words?

Exercise:

Problem: What does $P(E \text{ OR } M)$ mean in words?

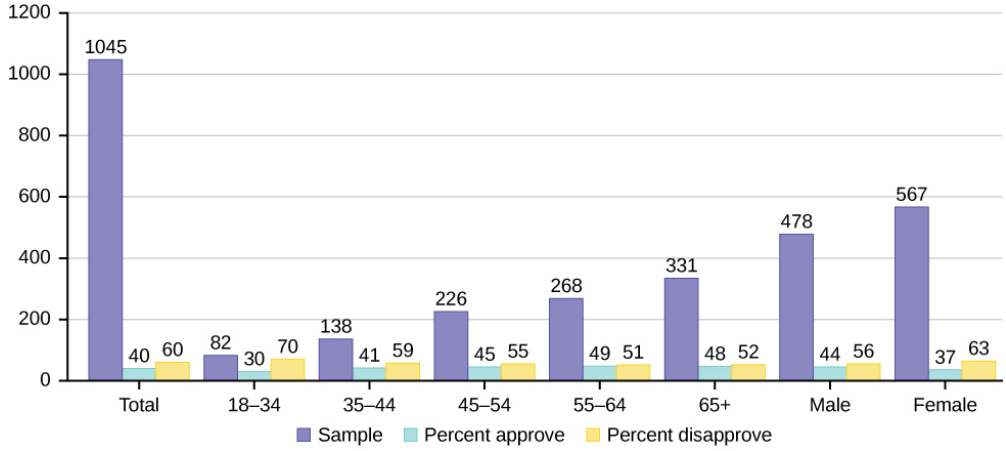
Solution:

the probability of landing on an even number or a multiple of three

Homework

Exercise:

Problem:



The graph in [link] displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

Exercise:

Problem:

Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

Solution:

a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

Glossary

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

Equally Likely

Each outcome of an experiment has the same probability.

Event

a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by *S*. An event is an arbitrary subset in *S*. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as *A*, *B*, *C*, and so on.

Experiment

a planned activity carried out under controlled conditions

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let *S* denote the sample space and *A* and *B* are two events in *S*. Then:

- $0 \leq P(A) \leq 1$
- If *A* and *B* are any two mutually exclusive events, then $P(A \text{ OR } B) = P(A) + P(\text{B})$.
- $P(S) = 1$

Sample Space

the set of all possible outcomes of an experiment

The Intersection: the AND Event

An outcome is in the event *A* AND *B* if the outcome is in both *A* AND *B* at the same time.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of A GIVEN B

$P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The Union: the OR Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events are independent if the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

Two events *A* and *B* are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether *A* and *B* are independent or dependent, **assume they are dependent until you can show otherwise**.

Example:
You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit.
a. Sampling with replacement:
Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the *Q* of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the *Q* of spades again. Your picks are {*Q* of spades, ten of clubs, *Q* of spades}. You have picked the *Q* of spades twice. You pick each card from the 52-card deck.
b. Sampling without replacement:
Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the *K* of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the *J* of spades. Your picks are {*K* of hearts, three of diamonds, *J* of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice. The probability of picking the three of diamonds is called a conditional probability because it is conditioned on what was picked first. This is true also of the probability of picking the J of spades. The probability of picking the J of spades is actually conditioned on *both* the previous picks.

Note:
Try It
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are *Q* of spades, *K* of hearts and *Q* of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are *Q* of spades, *K* of hearts, and *J* of spades. Can you decide if the sampling was with or without replacement?

Solution:

a. With replacement
b. No

Example:
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are *QS*, 1*D*, 1*C*, *QD*.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are *KH*, 7*D*, 6*D*, *KH*.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Solution:

a. Without replacement; b. With replacement

Note:
Try It
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. *QS*, 1*D*, 1*C*, *QD*
b. *KH*, 7*D*, 6*D*, *KH*
c. *QS*, 7*D*, 6*D*, *KS*

Solution:

without replacement: 1. Possible; 2. Impossible, 3. Possible

with replacement: 1. Possible; 2. Possible, 3. Possible

Mutually Exclusive Events

A and *B* are **mutually exclusive** events if they cannot occur at the same time. This means that *A* and *B* do not share any outcomes and *P*(*A* AND *B*) = 0.

For example, suppose the sample space *S* = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let *A* = {1, 2, 3, 4, 5}, *B* = {4, 5, 6, 7, 8}, and *C* = {7, 9}. *A* AND *B* = {4, 5}. *P*(*A* AND *B*) = — and is not equal to zero. Therefore, *A* and *B* are not mutually exclusive. *A* and *C* do not have any numbers in common so *P*(*A* AND *C*) = 0. Therefore, *A* and *C* are mutually exclusive.

If it is not known whether *A* and *B* are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example:
Exercise:

Problem: Flip two fair coins. Find the probabilities of the events.

a. Let *F* = the event of getting at most one tail (zero or one tail).
b. Let *G* = the event of getting two faces that are the same.
c. Let *H* = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are *F* and *G* mutually exclusive?
e. Let *J* = the event of getting all tails. Are *J* and *H* mutually exclusive?

Solution:

Look at the sample space in [link].

a. Zero (0) or one (1) tails occur when the outcomes *HH*, *TH*, *HT* show up. *P*(*F*) = —
b. Two faces are the same if *HH* or *TT* show up. *P*(*G*) = —

c. A head on the first flip followed by a head or tail on the second flip occurs when *HH* or *HT* show up. $P(H) = \frac{\quad}{\quad}$
d. *F* and *G* share *HH* so *P*(*F* AND *G*) is not equal to zero (0). *F* and *G* are not mutually exclusive.
e. Getting all tails occurs when tails shows up on both coins (*TT*). *H*'s outcomes are *HH* and *HT*.

J and *H* have nothing in common so *P*(*J* AND *H*) = 0. *J* and *H* are mutually exclusive.

Note:
Try It
Exercise:

Problem:

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let *F* = the event of getting the white ball twice.
b. Let *G* = the event of getting two balls of different colors.
c. Let *H* = the event of getting white on the first pick.
d. Are *F* and *G* mutually exclusive?
e. Are *G* and *H* mutually exclusive?

Solution:

a. $P(F) = \frac{\quad}{\quad}$
b. $P(G) = \frac{\quad}{\quad}$
c. $P(H) = \frac{\quad}{\quad}$
d. Yes
e. No

Example:
Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then A = {1, 3, 5}. Let event B = a face is even. Then B = {2, 4, 6}.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = -$ and $P(B) = -$.
- Let event C = odd faces larger than two. Then C = {3, 5}. Let event D = all even faces smaller than five. Then D = {2, 4}. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. E = {1, 2, 3, 4}.

Exercise:

Problem:

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Solution:

No. C = {3, 5} and E = {1, 2, 3, 4}. $P(C \text{ AND } E) = -$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = -$.

Note:
Try It

Exercise:

Problem:

Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(A|B) = P(A)$
- $P(B|A)$
- $P(A \text{ AND } B) = P(A)P(B)$

Solution:

$P(A|B) =$ ———— ——

The events are independent because $P(A|B) = P(A)$.

Example:
Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?
If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

Note:
NOTE
The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary

information.

Exercise:

Problem: a. Show that $P(G|H) = P(G)$.

Solution:

$P(G|H) = \frac{\text{————}}{\text{————}} = \frac{\text{—}}{\text{—}} = 0.6 = P(G)$

Exercise:

Problem: b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution:

$P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ AND } H)$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

Note:
Try It
Exercise:

Problem:

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(G \text{ AND } O)$?

Solution:

Event G and $O = \{G1, G3\}$

$P(G \text{ and } O) = \frac{2}{10} = 0.2$

Example:
Exercise:

Problem:

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(D|C)$?

Solution:

a. Yes, because $P(C|D) = P(C)$.
b. No, because $P(C \text{ AND } D)$ is not equal to zero.

c. $P(D|C) = \frac{\quad}{\quad} = \frac{\quad}{\quad} = 0.3$

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(B \text{ AND } D) = 0.20$.

a. Find $P(B|D)$.
b. Find $P(D|B)$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Solution:

a. $P(B|D) = 0.6667$
b. $P(D|B) = 0.5$
c. No
d. No

Example:
In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.
The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$. S has eight outcomes.

- $P(R) = -$. $P(B) = -$. $P(R \text{ AND } B) = 0$. (You cannot draw one card that is both red and blue.)
- $P(E) = -$. (There are three even-numbered cards, $R2$, $B2$, and $B4$.)
- $P(E|B) = -$. (There are five blue cards: $B1$, $B2$, $B3$, $B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(B|E) = -$. (There are three even-numbered cards: $R2$, $B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(R \text{ AND } B) = 0$.
- Let G = card with a number greater than 3. $G = \{B4, B5\}$. $P(G) = -$. Let H = blue card numbered between one and four, inclusive. $H = \{B1, B2, B3, B4\}$. $P(G|H) = -$. (The only card in H that has a number greater than three is $B4$.) Since $- = -$, $P(G) = P(G|H)$, which means that G and H are independent.

Note:
Try It
Exercise:

Problem: In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.
Let B be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Solution:

$P(B|A) = 0.67$

$P(B) = 0.25$

So $P(B)$ does not equal $P(B|A)$ which means that B and A are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\text{B AND } A) = 0.20$, not 0.

Example:
In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L|F) = 0.75$

Note:
NOTE
The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(F|L)$ yet, so you cannot use the second condition.

Solution 1
Check whether $P(F \text{ AND } L) = P(F)P(L)$. We are given that $P(F \text{ AND } L) = 0.45$, but $P(F)P(L) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(F \text{ AND } L)$ does not equal $P(F)P(L)$.
Solution 2
Check whether $P(L|F)$ equals $P(L)$. We are given that $P(L|F) = 0.75$, but $P(L) = 0.50$; they are not equal. The events of being female and having long hair are not independent.
Interpretation of Results
The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Note:
Try It
Exercise:

Problem:

Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = 0.44$ and $P(F) = 0.55$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

Solution:

Because $P(I \text{ AND } F) = 0$,

$P(I \text{ OR } F) = P(I) + P(F) - P(I \text{ AND } F) = 0.44 + 0.56 - 0 = 1$

Example:
Exercise:

Problem:

a. Toss one fair coin (the coin has two sides, *H* and *T*). The outcomes are _____. Count the outcomes. There are ____ outcomes.
b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are _____. Count the outcomes. There are ___ outcomes.
c. Multiply the two numbers of outcomes. The answer is _____.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in three is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are *H*1 and *T*6.)
e. Event *A* = heads (*H*) on the coin followed by an even number (2, 4, 6) on the die.
   A = {_____}. Find *P*(*A*).
f. Event *B* = heads on the coin followed by a three on the die. *B* = {_____}. Find *P*(*B*).
g. Are *A* and *B* mutually exclusive? (Hint: What is *P*(*A* AND *B*)? If *P*(*A* AND *B*) = 0, then *A* and *B* are mutually exclusive.)
h. Are *A* and *B* independent? (Hint: Is *P*(*A* AND *B*) = *P*(*A*)*P*(*B*)? If *P*(*A* AND *B*) = *P*(*A*)*P*(*B*), then *A* and *B* are independent. If not, then they are dependent).

Solution:

a. *H* and *T*; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. *T*1, *T*2, *T*3, *T*4, *T*5, *T*6, *H*1, *H*2, *H*3, *H*4, *H*5, *H*6
e. *A* = {*H*2, *H*4, *H*6}; *P*(*A*) = —
f. *B* = {*H*3}; *P*(*B*) = —
g. Yes, because *P*(*A* AND *B*) = 0

h. $P(A \text{ AND } B) = 0$. $P(A)P(B) =$ — — . $P(A \text{ AND } B)$ does not equal $P(A)P(B)$, so A and B are dependent.

Note:
Try It
Exercise:

Problem:

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let T be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

a. Compute $P(T)$.
b. Compute $P(T|F)$.
c. Are T and F independent?.
d. Are F and S mutually exclusive?
e. Are F and S independent?

Solution:

a. $P(T) =$ –
b. $P(T|F) =$ –
c. No
d. No
e. Yes

References

Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Chapter Review

Two events *A* and *B* are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

If *A* and *B* are independent, $P(A \cap B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

If *A* and *B* are mutually exclusive, $P(A \cup B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$.

Exercise:

Problem:

E and *F* are mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E|F)$.

Exercise:

Problem: *J* and *K* are independent events. $P(J|K) = 0.3$. Find $P(J)$.

Solution:

$P(J) = 0.3$

Exercise:

Problem:

U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

a. $P(U \text{ AND } V) =$
b. $P(U|V) =$
c. $P(U \text{ OR } V) =$

Exercise:

Problem:

Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ AND } R) = 0.1$. Find $P(R)$.

Solution:

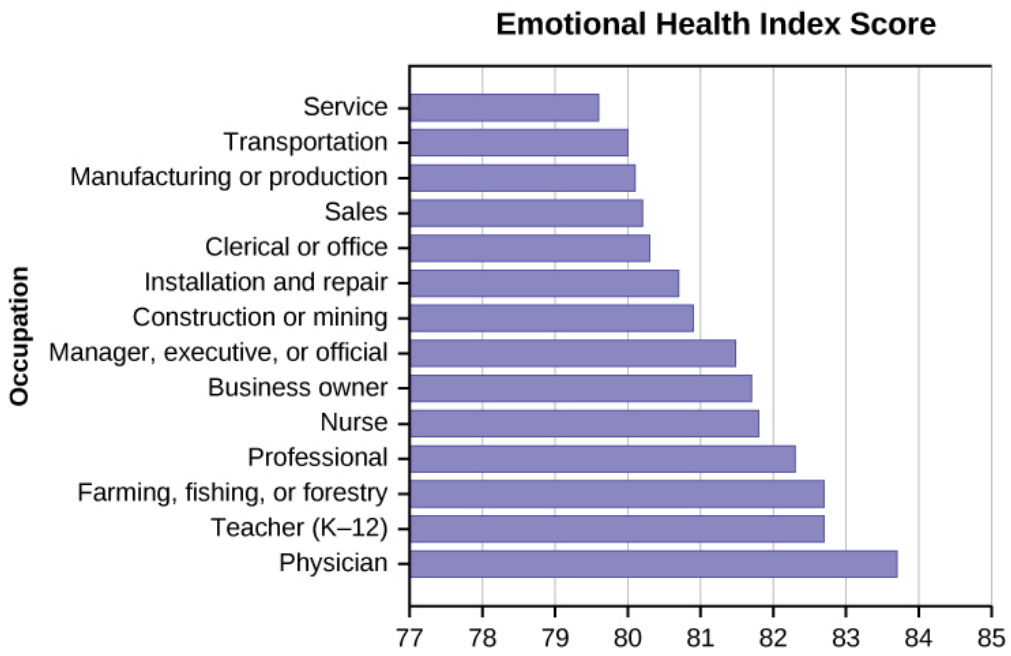
$P(Q \text{ AND } R) = P(Q)P(R)$

$0.1 = (0.4)P(R)$

$P(R) = 0.25$

Homework

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.



Exercise:

Problem:

Find the probability that an Emotional Health Index Score is 82.7.

Exercise:

Problem:

Find the probability that an Emotional Health Index Score is 81.0.

Solution:

0

Exercise:

Problem:

Find the probability that an Emotional Health Index Score is more than 81?

Exercise:

Problem:

Find the probability that an Emotional Health Index Score is between 80.5 and 82?

Solution:

0.3571

Exercise:

Problem:

If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

Exercise:

Problem:

What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

Solution:

0.2142

Exercise:

Problem:

What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

Exercise:

Problem: What occupation has the highest emotional index score?

Solution:

Physician (83.7)

Exercise:

Problem: What occupation has the lowest emotional index score?

Exercise:

Problem: What is the range of the data?

Solution:

$83.7 - 79.6 = 4.1$

Exercise:

Problem: Compute the average EHIS.

Exercise:

Problem:

If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

Solution:

$P(\text{Occupation} < 81.3) = 0.5$

Bringing It Together

Exercise:

Problem:

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into [link].

Shirt#	≤ 210	211–250	251–290	290≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about P(Shirt# 1–33|≤ 210 pounds)?

Exercise:

Problem:

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

a. $P(C)$ = _____
b. $P(P|C)$ = _____
c. $P(P|C')$ = _____
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Solution:

a. $P(C) = 0.4567$

b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise:

Problem:

Given events G and H: $P(G) = 0.43$; $P(H) = 0.26$; $P(H \text{ AND } G) = 0.14$

a. Find $P(H \text{ OR } G)$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

Exercise:

Problem:

Given events J and K: $P(J) = 0.18$; $P(K) = 0.37$; $P(J \text{ OR } K) = 0.45$

a. Find $P(J \text{ AND } K)$.
b. Find the probability of the complement of event (J AND K).
c. Find the probability of the complement of event (J AND K).

Solution:

a. $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K)$; $0.45 = 0.18 + 0.37 - P(J \text{ AND } K)$; solve to find $P(J \text{ AND } K) = 0.10$
b. $P(\text{NOT } (J \text{ AND } K)) = 1 - P(J \text{ AND } K) = 1 - 0.10 = 0.90$
c. $P(\text{NOT } (J \text{ OR } K)) = 1 - P(J \text{ OR } K) = 1 - 0.45 = 0.55$

Glossary

Dependent Events

If two events are NOT independent, then we say that they are dependent.

Sampling with Replacement

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement

When sampling is done without replacement, each member of a population may be chosen only once.

Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If *A* and *B* are two events defined on a **sample space**, then: $P(A \cap B) = P(B)P(A|B)$. We can think of the intersection symbol as substituting for the word "and".

This rule may also be written as: $P\left(A\middle|B\right) = \frac{P(A \cap B)}{P(B)}$

This equation is read as the probability of *A* given *B* equals the probability of *A* and *B* divided by the probability of *B*.

If *A* and *B* are **independent**, then $P(A|B) = P(A)$. Then $P(A \cap B) = P(A|B)P(B)$ becomes $P(A \cap B) = P(A)(B)$ because the $P(A|B) = P(A)$ if *A* and *B* are independent.

One easy way to remember the multiplication rule is that the word "and" means that the event has to satisfy two conditions. For example the name drawn from the class roster is to be both a female and a sophomore. It is harder to satisfy two conditions than only one and of course when we multiply fractions the result is always smaller. This reflects the increasing difficulty of satisfying two conditions.

The Addition Rule

If *A* and *B* are defined on a sample space, then: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can think of the union symbol substituting for the word "or". The reason we subtract the intersection of *A* and *B* is to keep from double counting elements that are in both *A* and *B*.

If *A* and *B* are **mutually exclusive**, then $P(A \cap B) = 0$. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes

$P(A \cup B) = P(A) + P(B)$.

Example:
Klaus is trying to choose where to go on vacation. His two choices are: *A* = New Zealand and *B* = Alaska

- Klaus can only afford one vacation. The probability that he chooses *A* is *P*(*A*) = 0.6 and the probability that he chooses *B* is *P*(*B*) = 0.35.
- $P(A \cap B) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(A \cup B) = P(A) + P(B) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Example:
Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. *A* = the event Carlos is successful on his first attempt. *P*(*A*) = 0.65. *B* = the event Carlos is successful on his second attempt. *P*(*B*) = 0.65. Carlos tends to shoot in streaks. The probability that he makes the second goal | that he made the first goal is 0.90.

Exercise:

Problem: a. What is the probability that he makes both goals?

Solution:

a. The problem is asking you to find $P(A \cap B) = P(B \cap A)$. Since *P*(*B*|*A*) = 0.90: *P*(*B* ∩ *A*) = *P*(*B*|*A*) *P*(*A*) = (0.90)(0.65) = 0.585

Carlos makes the first and second goals with probability 0.585.

Exercise:

Problem:

b. What is the probability that Carlos makes either the first goal or the second goal?

Solution:

b. The problem is asking you to find $P(A \cup B)$.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.65 + 0.65 - 0.585 = 0.715$

Carlos makes either the first goal or the second goal with probability 0.715.

Exercise:

Problem: c. Are A and B independent?

Solution:

c. No, they are not, because $P(B \cap A) = 0.585$.

$P(B)P(A) = (0.65)(0.65) = 0.423$

$0.423 \neq 0.585 = P(B \cap A)$

So, $P(B \cap A)$ is **not** equal to $P(B)P(A)$.

Exercise:

Problem: d. Are A and B mutually exclusive?

Solution:

d. No, they are not because $P(A \cap B) = 0.585$.

To be mutually exclusive, $P(A \cap B)$ must equal zero.

Note:
Try It
Exercise:

Problem:

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(C) = 0.75$. D = the event Helen makes the second shot. $P(D) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

Solution:

$P(D|C) = 0.85$

$P(C \cap D) = P(D \cap C)$
$P(D \cap C) = P(D|C)P(C) = (0.85)(0.75) = 0.6375$
Helen makes the first and second free throws with probability 0.6375.

Example:
A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

Exercise:

Problem:

a. What is the probability that the member is a novice swimmer?

Solution:

a. $\frac{28}{150}$

Exercise:

Problem:

b. What is the probability that the member practices four times a week?

Solution:

b. $\frac{80}{150}$

Exercise:

Problem:

c. What is the probability that the member is an advanced swimmer and practices four times a week?

Solution:

c. $\frac{40}{150}$

Exercise:

Problem:

d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

Solution:

d. $P(\text{advanced} \cap \text{intermediate}) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

Exercise:

Problem:

e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Solution:

e. No, these are not independent events.
$P(\text{novice} \cap \text{practices four times per week}) = 0.0667$
$P(\text{novice})P(\text{practices four times per week}) = 0.0996$
$0.0667 \neq 0.0996$

Note:
Try It
Exercise:

Problem:

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Solution:

$P = \frac{200-140-40}{200} = \frac{20}{200} = 0.1$

Example:
Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class | that she enrolls in speech class is 0.25.
Let: M = math class, S = speech class, $M|S$ = math given speech

Exercise:

Problem:

a. What is the probability that Felicity enrolls in math and speech? Find $P(M \cap S) = P(M|S)P(S)$.
b. What is the probability that Felicity enrolls in math or speech classes?
Find $P(M \cup S) = P(M) + P(S) - P(M \cap S)$.
c. Are M and S independent? Is $P(M|S) = P(M)$?
d. Are M and S mutually exclusive? Is $P(M \cap S) = 0$?

Solution:

a. 0.1625, b. 0.6875, c. No, d. No

Note:
Try It

Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

a. Find $P(B \cap D)$.
b. Find $P(B \cup D)$.

Solution:

a. $P(B \cap D) = P(D|B)P(B) = (0.5)(0.4) = 0.20$.
b. $P(B \cup D) = P(B) + P(D) - P(B \cap D) = 0.40 + 0.30 - 0.20 = 0.50$

Example:
Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

Exercise:

Problem:

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

Solution:

a. $P(B) = 0.143$; $P(N) = 0.85$

Exercise:

Problem:

b. Given that the woman has breast cancer, what is the probability that she tests negative?

Solution:

b. $P(N|B) = 0.02$

Exercise:

Problem:

c. What is the probability that the woman has breast cancer AND tests negative?

Solution:

c. $P(B \cap N) = P(B)P(N|B) = (0.143)(0.02) = 0.0029$

Exercise:

Problem:

d. What is the probability that the woman has breast cancer or tests negative?

Solution:

d. $P(B \cup N) = P(B) + P(N) - P(B \cap N) = 0.143 + 0.85 - 0.0029 = 0.9901$

Exercise:

Problem:

e. Are having breast cancer and testing negative independent events?

Solution:

e. No. $P(N) = 0.85$; $P(N|B) = 0.02$. So, $P(N|B)$ does not equal $P(N)$.

Exercise:

Problem:

f. Are having breast cancer and testing negative mutually exclusive?

Solution:

f. No. $P(B \cap N) = 0.0029$. For B and N to be mutually exclusive, $P(B \cap N)$ must be zero.

Note:
Try It
Exercise:

Problem:

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Solution:

Let A = student is a senior going to college.

Let B = student plays sports.

$P(B) = \frac{140}{200}$

$P(B|A) = \frac{50}{140}$

$P(A \cap B) = P(B|A)P(A)$

$P(A \cap B) = \left(\frac{140}{200}\right)\left(\frac{50}{140}\right) = \frac{1}{4}$

Example:
Exercise:

Problem: Refer to the information in [link]. P = tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(P|B) = 1 - P(N|B)$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(B \cap P) = P(P|B)P(B)$.
c. What is the probability that a woman does not develop breast cancer. Find $P(B') = 1 - P(B)$.
d. What is the probability that a woman tests positive for breast cancer. Find $P(P) = 1 - P(N)$.

Solution:

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

a. Find $P(B')$.
b. Find $P(D \cap B)$.
c. Find $P(B|D)$.
d. Find $P(D \cap B')$.
e. Find $P(D|B')$.

Solution:

a. $P(B') = 0.60$
b. $P(D \cap B) = P(D|B)P(B) = 0.20$
c. $P(B|D) = \frac{P(B \cap D)}{P(D)} = \frac{(0.20)}{(0.30)} = 0.66$

d. $P(D \cap B') = P(D) - P(D \cap B) = 0.30 - 0.20 = 0.10$
e. $P(D|B') = P(D \cap B')P(B') = (P(D) - P(D \cap B))(0.60) = (0.10)(0.60) = 0.06$

References

DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf (accessed May 2, 2013).

Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

“Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

“Roulette.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at http://www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Chapter Review

The multiplication rule and the addition rule are used for computing the probability of *A* and *B*, as well as the probability of *A* or *B* for two given events *A*, *B* defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events *A* and *B* are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(A \cap B) = P(A|B)P(B)$

The addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.

Exercise:

Problem: Find $P(C)$.

Exercise:

Problem: Find $P(L)$.

Solution:

0.376

Exercise:

Problem: Find $P(C|L)$.

Exercise:

Problem: In words, what is $C|L$?

Solution:

$C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Exercise:

Problem: Find $P(L \cap C)$.

Exercise:

Problem: In words, what is $L \cap C$?

Solution:

$L \cap C$ is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty

for a person convicted of first degree murder.

Exercise:

Problem: Are L and C independent events? Show why or why not.

Exercise:

Problem: Find $P(L \cup C)$.

Solution:

0.6492

Exercise:

Problem: In words, what is $L \cup C$?

Exercise:

Problem:

Are L and C mutually exclusive events? Show why or why not.

Solution:

No, because $P(L \cap C)$ does not equal 0.

Homework

Exercise:

Problem:

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- C = California registered voters who support same-sex marriage.
- B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- A = California registered voters who are 18 to 39 years old.

a. Find $P(C)$.
b. Find $P(B)$.
c. Find $P(C|A)$.
d. Find $P(B|C)$.
e. In words, what is $C|A$?
f. In words, what is $B|C$?
g. Find $P(C \cap B)$.
h. In words, what is $C \cap B$?
i. Find $P(C \cup B)$.
j. Are C and B mutually exclusive events? Show why or why not.

Exercise:

Problem:

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

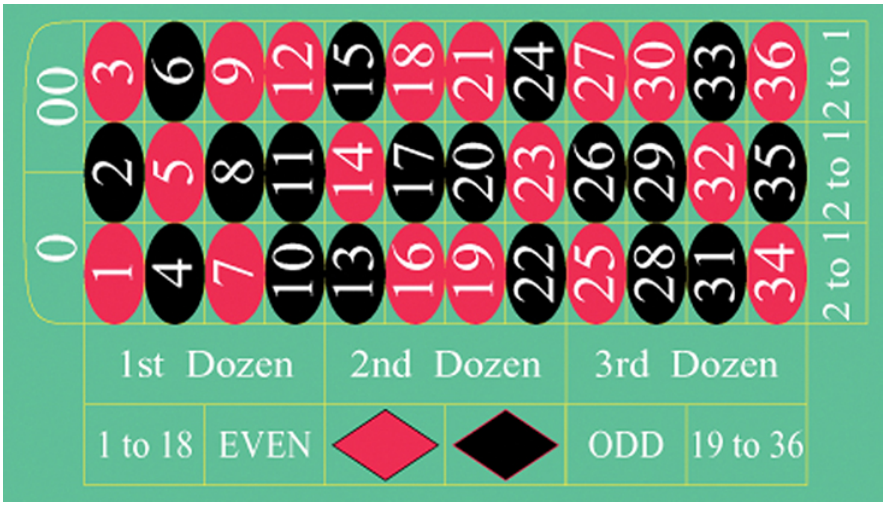
- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

Solution:

a. The Forum Research surveyed 1,046 Torontonians.
b. 58%
c. 42% of 1,046 = 439 (rounding to the nearest integer)
d. 0.57
e. 0.60.

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.



(credit: film8ker/wikibooks)

Exercise:

Problem:

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find *P*(red).
c. You bet on -1st 12- (1st Dozen). Find *P*(-1st 12-).
d. You bet on an even number. Find *P*(even number).
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

Exercise:

Problem:

Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number

d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

Solution:

a. P(Betting on two line that touch each other on the table) = $\frac{6}{38}$
b. P(Betting on three numbers in a line) = $\frac{3}{38}$
c. P(Bettting on one number) = $\frac{1}{38}$
d. P(Betting on four number that touch each other to form a square) = $\frac{4}{38}$
e. P(Betting on two number that touch each other on the table) = $\frac{2}{38}$
f. P(Betting on 0-00-1-2-3) = $\frac{5}{38}$
g. P(Betting on 0-1-2; or 0-00-2; or 00-2-3) = $\frac{3}{38}$

Exercise:

Problem:

Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

Exercise:

Problem:

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

a. List the sample space.
b. $P(G) =$ _____
c. $P(G|E) =$ _____
d. $P(G \cap E) =$ _____
e. $P(G \cup E) =$ _____
f. Are G and E mutually exclusive? Justify your answer numerically.

Solution:

a. {$G1$, $G2$, $G3$, $G4$, $G5$, $Y1$, $Y2$, $Y3$}
b. $\frac{5}{8}$
c. $\frac{2}{3}$
d. $\frac{2}{8}$
e. $\frac{6}{8}$
f. No, because $P(G \cap E)$ does not equal 0.

Exercise:

Problem: Roll two fair dice separately. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
c. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.

d. In words, explain what "$P(A|B)$" represents. Find $P(A|B)$.
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

Exercise:

Problem:

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Solution:

Note:
NOTE
The coin toss is independent of the card picked first.

a. $\{(G,H)\ (G,T)\ (B,H)\ (B,T)\ (R,H)\ (R,T)\}$
b. $P(A) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$

c. Yes, A and B are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \cap B) = 0$
d. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(A \cap C) = P(A) = \frac{3}{20}$

Exercise:

Problem:

An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Exercise:

Problem:

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find $P(A)$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

Solution:

a. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$
b. $\frac{4}{8}$

c. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(A \cap B) = 0$.

Exercise:

Consider the following scenario:
Let $P(C) = 0.4$.
Let $P(D) = 0.5$.
Problem: Let $P(C|D) = 0.6$.

a. Find $P(C \cap D)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(C \cup D)$.
e. Find $P(D|C)$.

Exercise:

Problem: Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z)$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(Z)$ if $P(Y \cup Z) = 0.71$ and $P(Y) = 0.42$.

Solution:

a. If Y and Z are independent, then $P(Y \cap Z) = P(Y)P(Z)$, so $P(Y \cup Z) = P(Y) + P(Z) - P(Y)P(Z)$.
b. 0.5

Exercise:

Problem: G and H are mutually exclusive events. $P(G) = 0.5$ $P(H) = 0.3$

a. Explain why the following statement MUST be false: $P(H|G) = 0.4$.

b. Find $P(H \cup G)$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

Exercise:

Problem:

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: E = speaks English at home; E' = speaks another language at home; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	**Answers**
a. $P(E')$ =	i. 0.8043
b. $P(E)$ =	ii. 0.623
c. $P(S \cap E')$ =	iii. 0.1957
d. $P(S\|E')$ =	iv. 0.1219

Solution:

iii i iv ii

Exercise:

Problem:

1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let *G* = won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let *F* = was a finalist.
c. Are *G* and *F* independent or dependent events? Justify your answer numerically and also explain why.
d. Are *G* and *F* mutually exclusive events? Justify your answer numerically and explain why.

Exercise:

Problem:

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: *R* = money returned; *E* = economics classes; *O* = other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.

c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

Solution:

a. $P(R) = 0.44$
b. $P(R|E) = 0.56$
c. $P(R|O) = 0.31$
d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R|E) \neq P(R)$.
e. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

Exercise:

Problem:

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home run	Total hits

Name	Single	Double	Triple	Home run	Total hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because $P(\text{hit by Hank Aaron} \mid \text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron} \mid \text{hit is a double}) \neq P(\text{hit is a double})$
c. No, because $P(\text{hit is by Hank Aaron} \mid \text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
d. Yes, because $P(\text{hit is by Hank Aaron} \mid \text{hit is a double}) = P(\text{hit is a double})$

Exercise:

Problem:

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

Solution:

a. $P(\text{type O} \cup \text{Rh-}) = P(\text{type O}) + P(\text{Rh-}) - P(\text{type O} \cap \text{Rh-})$

   $0.52 = 0.43 + 0.15 - P(\text{type O} \cap \text{Rh-})$; solve to find $P(\text{type O} \cap \text{Rh-}) = 0.06$

   6% of people have type O, Rh- blood

b. $P(\text{NOT}(\text{type O} \cap \text{Rh-})) = 1 - P(\text{type O} \cap \text{Rh-}) = 1 - 0.06 = 0.94$

   94% of people do not have type O, Rh- blood

Exercise:

Problem:

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

a. Find the probability that a course has a final exam or a research project.

b. Find the probability that a course has NEITHER of these two requirements.

Exercise:

Problem:

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

a. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

Solution:

a. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
b. $P(C \cup N) = P(C) + P(N) - P(C \cap N) = 0.36 + 0.12 - 0.08 = 0.40$
c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(C \cup N) = 1 - 0.40 = 0.60$

Exercise:

Problem:

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part time student

a. Find $P(D \cap E)$.
b. Find $P(E|D)$.
c. Find $P(D \cup E)$.
d. Using an appropriate test, show whether D and E are independent.

e. Using an appropriate test, show whether D and E are mutually exclusive.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A \cap B) = 0$.

Contingency Tables and Tree Diagrams -- Probability Topics -- MtRoyal - Version2016RevA

Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example:
Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.
Calculate the following probabilities using the table.

Exercise:

Problem: a. Find *P*(Person is a car phone user).

Solution:

a. $\frac{\text{number of car phone users}}{\text{total number in study}} = \frac{305}{755}$

Exercise:

Problem: b. Find *P*(person had no violation in the last year).

Solution:

b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$

Exercise:

Problem:

c. Find *P*(Person had no violation in the last year AND was a car phone user).

Solution:

c. $\frac{280}{755}$

Exercise:

Problem:

d. Find *P*(Person is a car phone user OR person had no violation in the last year).

Solution:

d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$

Exercise:

Problem:

e. Find *P*(Person is a car phone user GIVEN person had a violation in the last year).

Solution:

e. $\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)

Exercise:

Problem:

f. Find *P*(Person had no violation last year GIVEN person was not a car phone user)

Solution:

f. $\frac{405}{450}$ (The sample space is reduced to the number of persons who were not car phone users.)

Note:
Try it

Exercise:

Problem:

[link] shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	**No injury in last year**	**Total**
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. What is P(athlete stretches before exercising)?
b. What is P(athlete stretches before exercising|no injury in the last year)?

Solution:

a. P(athlete stretches before exercising) = $\frac{350}{800}$ = 0.4375
b. P(athlete stretches before exercising|no injury in the last year) = $\frac{295}{514}$ = 0.5739

Example:

[link] shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Hiking Area Preference

Exercise:

Problem: a. Complete the table.

Solution:

a.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

Hiking Area Preference

Exercise:

Problem:

b. Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

Solution:

b.

1. $P(F \text{ AND } C) = \frac{18}{100} = 0.18$
2. $P(F)P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$

$P(F \text{ AND } C) \neq P(F)P(C)$, so the events F and C are not independent.

Exercise:

Problem:

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: $P(___|___) = ___$.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

Solution:

c.

1. The word 'given' tells you that this is a conditional.
2. $P(M|L) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

Exercise:

Problem:

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = prefers mountain peaks.

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

Solution:

d.

1. $P(F) = \frac{45}{100}$
2. $P(P) = \frac{25}{100}$
3. $P(F \text{ AND } P) = \frac{11}{100}$
4. $P(F \text{ OR } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Note:
Try It
Exercise:

Problem:

[link] shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Solution:

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$

$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$

$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example:
Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Caught or Not	**Door One**	**Door Two**	**Door Three**	**Total**

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

Door Choice

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is *P*(Door One AND Caught)
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is *P*(Door One AND Not Caught)

Verify the remaining entries.

Exercise:

Problem:

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Solution:

a.

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$

Caught or Not	Door One	Door Two	Door Three	Total
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

Door Choice

Exercise:

Problem:

b. What is the probability that Alissa does not catch Muddy?

Solution:

b. $\frac{41}{60}$

Exercise:

Problem:

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution:

c. $\frac{9}{19}$

Example:
[link] contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Exercise:

Problem: TOTAL each column and each row. Total data = 4,520.7

a. Find *P*(2009 AND Robbery).
b. Find *P*(2010 AND Burglary).
c. Find *P*(2010 OR Burglary).
d. Find *P*(2011|Rape).
e. Find *P*(Vehicle|2008).

Solution:

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Note:
Try It
Exercise:

Problem:

[link] relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Solution:

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. $P(\text{Tall}) = \frac{50}{205} = 0.244$
c. $P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$
d. $P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$
e. $P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$
f. $P(\text{Tall AND Underweight} = \frac{12}{205} = 0.0585$
g. No. $P(\text{Tall})$ does not equal $P(\text{Tall}|\text{Obese})$.

Tree Diagrams

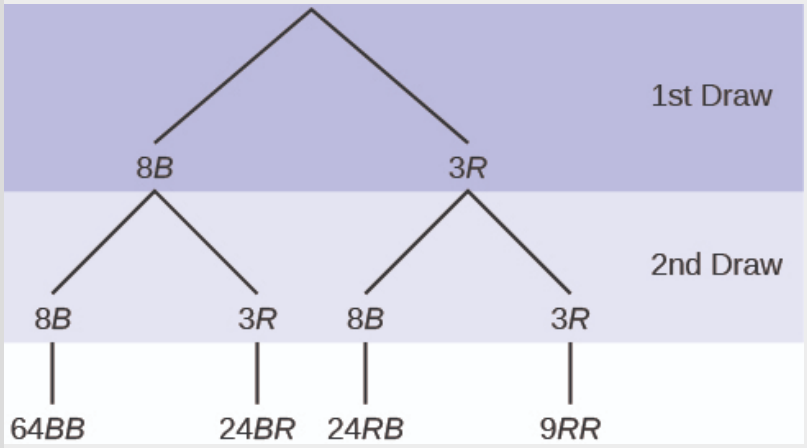
Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams can be used to visualize and solve conditional probabilities.

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability

problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example:
In an urn, there are 11 balls. Three balls are red (*R*) and eight balls are blue (*B*). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.



Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as *R*1, *R*2, and *R*3 and each blue ball as *B*1, *B*2, *B*3, *B*4, *B*5, *B*6, *B*7, and *B*8. Then the nine *RR* outcomes can be written as:
*R*1*R*1 *R*1*R*2 *R*1*R*3 *R*2*R*1 *R*2*R*2 *R*2*R*3 *R*3*R*1 *R*3*R*2 *R*3*R*3
The other outcomes are similar.
There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the **sample space**.

Exercise:

Problem: a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...

Solution:

a. *B*1*R*1 *B*1*R*2 *B*1*R*3 *B*2*R*1 *B*2*R*2 *B*2*R*3 *B*3*R*1 *B*3*R*2 *B*3*R*3 *B*4*R*1 *B*4*R*2 *B*4*R*3 *B*5*R*1 *B*5*R*2 *B*5*R*3 *B*6*R*1 *B*6*R*2 *B*6*R*3 *B*7*R*1 *B*7*R*2 *B*7*R*3 *B*8*R*1 *B*8*R*2 *B*8*R*3

Exercise:

Problem: b. Using the tree diagram, calculate *P*(*RR*).

Solution:

b. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

Exercise:

Problem: c. Using the tree diagram, calculate *P*(*RB* OR *BR*).

Solution:

c. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

Exercise:

Problem:

d. Using the tree diagram, calculate *P*(*R* on 1st draw AND *B* on 2nd draw).

Solution:

d. $P(R \text{ on 1st draw AND } B \text{ on 2nd draw}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

Exercise:

Problem:

e. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.

Solution:

e. $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw}) = P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{24}{88} = \frac{3}{11}$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 *BR* and 64 *BB*). Twenty-four of the 88 possible outcomes are *BR*. $\frac{24}{88} = \frac{3}{11}$.

Exercise:

Problem: f. Using the tree diagram, calculate $P(BB)$.

Solution:

f. $P(BB) = \frac{64}{121}$

Exercise:

Problem:

g. Using the tree diagram, calculate $P(B \text{ on the 2nd draw given } R \text{ on the first draw})$.

Solution:

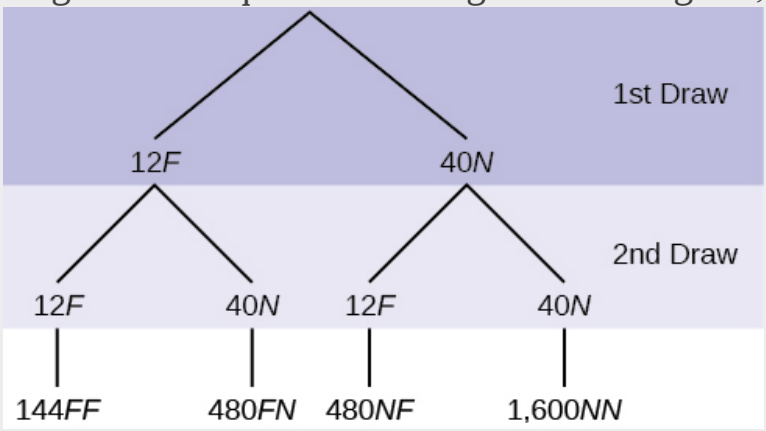
g. $P(B \text{ on 2nd draw} | R \text{ on 1st draw}) = \frac{8}{11}$

There are 9 + 24 outcomes that have *R* on the first draw (9 *RR* and 24 *RB*). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have *B* on the second draw. The probability is then $\frac{24}{33}$.

Note:
Try It
Exercise:

Problem:

In a standard deck, there are 52 cards. 12 cards are face cards (event *F*) and 40 cards are not face cards (event *N*). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate *P*(*FF*).

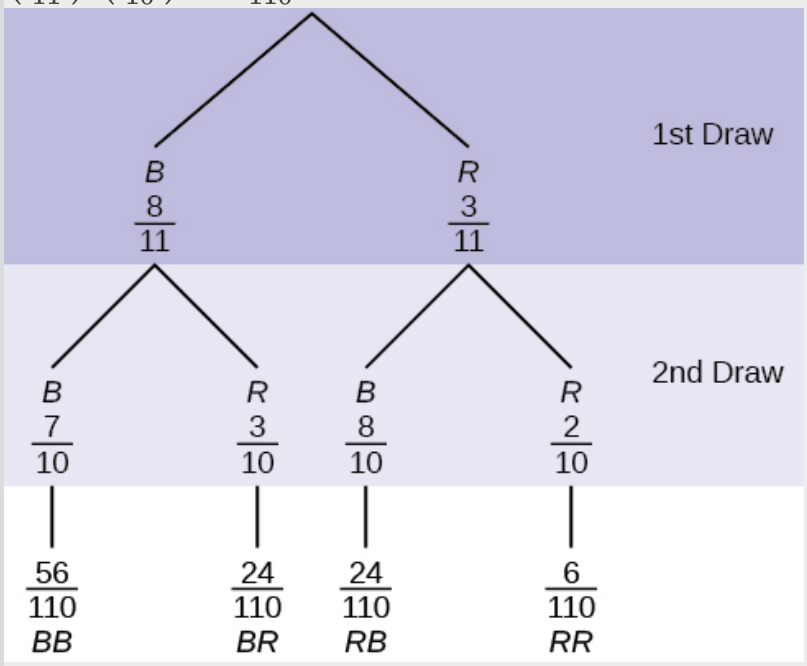


Solution:

Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$P(FF) = \frac{144}{144 + 480 + 480 + 1{,}600} = \frac{144}{2{,}704} = \frac{9}{169}$

Example:
An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.



$$\text{Total} = \frac{56+24+24+6}{110} = \frac{110}{110} = 1$$

Note:
NOTE
If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw **without**

replacement, so that on the second draw there are ten marbles left in the urn.

Calculate the following probabilities using the tree diagram.

Exercise:

Problem: a. $P(RR)$ = _____

Solution:

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

Exercise:

Problem: b. Fill in the blanks:

$P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (___)(___) = \frac{48}{110}$

Solution:

b. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

Exercise:

Problem: c. $P(R \text{ on 2nd} | B \text{ on 1st}) =$

Solution:

c. $P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{3}{10}$

Exercise:

Problem: d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (___)(___) = \frac{24}{100}$

Solution:

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$

Exercise:

Problem: e. Find $P(BB)$.

Solution:

e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

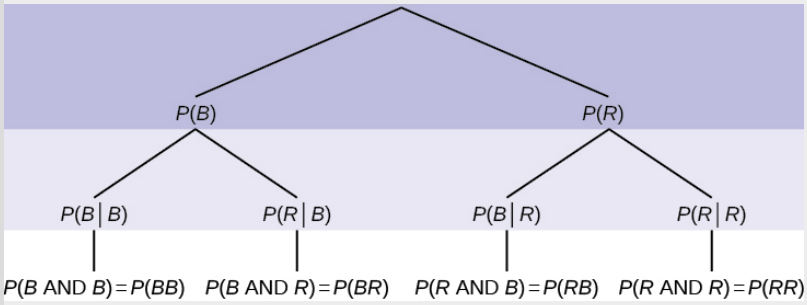
Exercise:

Problem: f. Find $P(B \text{ on 2nd}|R \text{ on 1st})$.

Solution:

f. Using the tree diagram, $P(B \text{ on 2nd}|R \text{ on 1st}) = P(R|B) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.



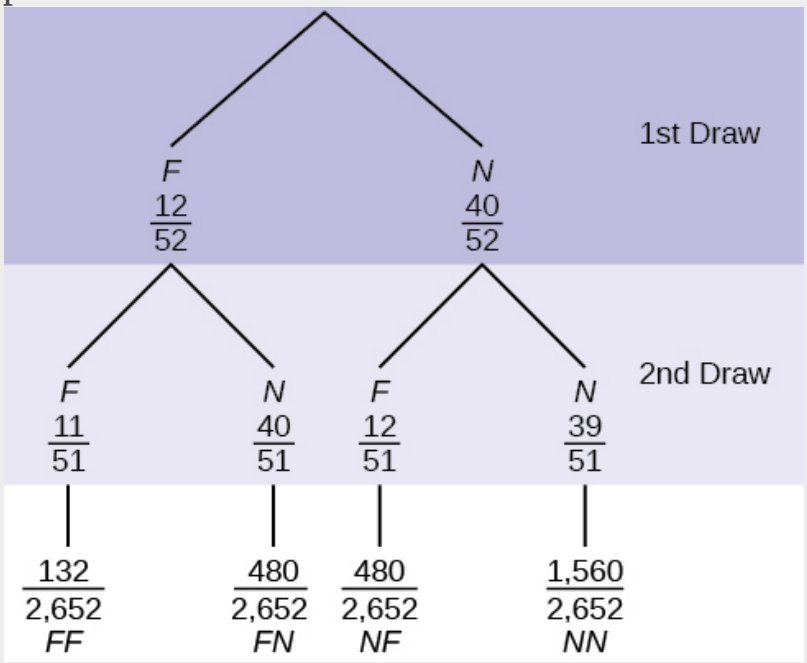
- $P(R|R)$ here means $P(R \text{ on 2nd}|R \text{ on 1st})$

- $P(B|R)$ here means $P(B$ on 2nd$|R$ on 1st$)$
- $P(R|B)$ here means $P(R$ on 2nd$|B$ on 1st$)$
- $P(B|B)$ here means $P(B$ on 2nd$|B$ on 1st$)$

Note:
Try It
Exercise:

Problem:

In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.



a. Find $P(FN$ OR $NF)$.
b. Find $P(N|F)$.
c. Find P(at most one face card).
   Hint: "At most one face card" means zero or one face card.
d. Find P(at least on face card).
   Hint: "At least one face card" means one or two face cards.

Solution:

a. $P(FN \text{ OR } NF) = \frac{480}{2{,}652} + \frac{480}{2{,}652} = \frac{960}{2{,}652} = \frac{80}{221}$

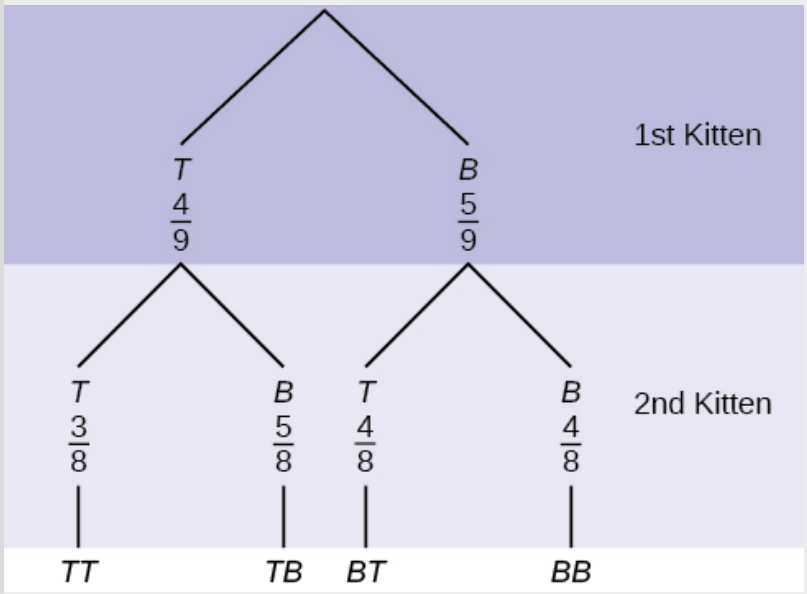
b. $P(N|F) = \frac{40}{51}$

c. $P(\text{at most one face card}) = \frac{(480 + 480 + 1{,}560)}{2{,}652} = \frac{2{,}520}{2{,}652}$

d. $P(\text{at least one face card}) = \frac{(132 + 480 + 480)}{2{,}652} = \frac{1{,}092}{2{,}652}$

Example:

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



Exercise:

Problem:

a. What is the probability that both kittens are tabby?

a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$

b. What is the probability that one kitten of each coloring is selected?

a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$

c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?

d. What is the probability of choosing two kittens of the same color?

Solution:

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Note:
Try It
Exercise:

Problem:

Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

Solution:

$\left(\frac{4}{7}\right)\left(\frac{3}{6}\right)+\left(\frac{3}{7}\right)\left(\frac{4}{6}\right)$

References

“Blood Types.” American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-about-blood/blood-types (accessed May

3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

“Human Blood Types.” Unite Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loīc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

“United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

Chapter Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilites that have multiple dependent variables.

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

Glossary

Tree Diagram
: the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Contingency Table
: the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

Introduction to Discrete Probability Distributions - MRU - C Lemieux (2017)

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to:

- Recognize the binomial probability distribution and apply it appropriately.
- Be able to evaluate evidence using the binomial distribution.

Suppose you flip a coin ten times and each time it comes up heads. This might make you start to wonder if there is something wrong with the coin. Perhaps it is a trick coin and is heads on both sides? Perhaps it is imbalanced and it is more likely to come up heads over tails. You may also wonder what is the probability of getting 10 heads in a row, if the coin was fair.

Coin flipping is interesting because it is a random event. We cannot predict whether the next flip will be heads or tails (assuming it isn't a trick coin). That means the outcome would be a random variable. A **random variable** is any variable where the outcome is determined by a random event. The outcome is also discrete because we count it. Above you flipped a coin ten times and counted the number of heads. A **discrete random variable** is a variable whose outcome is determined by a random event and where we count the outcomes. Other examples of discrete random variables include how many times you roll an even number with a die out of ten rolls; how many customers enter a store during a five-minute interval; how many times you draw a high card out of a deck of cards out of eight draws (without replacement).

In each of these situations (coin toss, rolling die, number of customers, drawing cards), you could look at each situation and, each time, come up with a new formula to find the probability of these events happening. But this would take a lot of work and be inefficient. Instead, you would want to

see if the situation can be **modelled** by a distribution. A **probability distribution** provides the theoretical probabilities of all of the possible events in a situation. For example, the following is a probability model of how many heads you can get when you flip a fair coin three times:

# of heads	Probability
0	0.125
1	0.375
2	0.375
3	0.125

Notice that the probabilities range from 0 to 1 and that the sum of the probabilities is 1.

The above table could be determined by working out all of the possible outcomes (TTT, TTH, THT, HTT, etc.), then counting how many heads were in each outcome. But again, that is time consuming. Instead, you want to see if there is a probability distribution that models your situation that you can use. For example, coin tossing can be modelled by the T distribution.

The binomial distribution is an example of a model for discrete random variables. There are many other models for discrete random variables including Poisson, geometric, hypergeometric, and discrete uniform to name a few. Each distribution comes with a set of criteria and if a situation fits that criteria, then the distribution can model it. That is, the distribution can produce theoretical probabilities for that situation.

Note:
Theoretical vs. experimental probabilities
In simplest terms, a theoretical probability is determined by using a formula while an experimental probability is found by actually doing the event. For example, if you flip a coin 3 times and get 2 heads, then the experimental probability is 2/3 = 0.6667. The theoretical probability is 0.375. The theoretical probability is also called the **long-run probability**, because the longer you do the experimental probability the closer the experimental results will get to the theoretical probability. This is an example of the **law of large numbers**.

In this chapter, we are going to learn about the binomial distribution, which is a model for discrete random variables. In the next chapter, we will learn about the normal distribution, which is a model for continuous random variables.

In particular, we want to use the binomial distribution to evaluate evidence. For example, going back to the example flipping a coin ten times and getting ten heads, we want to use the evidence (getting ten heads) to determine whether we think there is something wrong with the coin.

Binomial distribution - MRU - C Lemieux
An introduction to the binomial distribution an informal inferential statistics.

Binomial distribution

Flipping a coin a certain number of times, let's say ten times, is a classic example of a binomial distribution. What are the characteristics of flipping a coin that makes it binomial?

Before we answer that question, let's get a bit terminology out of the way. In probability theory, an **experiment** is the actual process that you investigating. In the flipping coin example, the experiment is flipping a coin ten times. A **trial** is a specific instance of an experiment. Flipping a coin only once is considered a trial.

Going back to the coin flipping example, let's assume that we are dealing with a fair coin (i.e. probability of getting a head or a tail is 50%). We've already discussed that when we count the number of heads that this is an example of a discrete random variable. Notice that there are only two possible outcomes (heads or tails). This is a key criterion for a binomial distribution (*binomial* derives from Latin for two terms). Also, notice that the events are independent of each other. That is, if you get a head on one flip, that has no impact on the probability of getting a head on the next flip. This also means that the probability of getting a head remains constant. This is another key criterion for a binomial distribution. The other thing to notice is that the number of times we flip the coin is fixed. We don't flip it until we get bored or run out of time. Instead we flip it ten times. This means that the number of trials is fixed. This is the last key criterion of a binomial distribution.

There are five characteristics of a binomial experiment.

1. The variable being studied is random.
2. The outcomes of the variable are being counted.
3. There are a fixed number of trials. The letter n denotes the number of trials.
4. There are only two possible outcomes, called "success" and "failure," for each trialπ denotes the probability of a success on one trial, and 1-π denotes the probability of a failure on one trial.
5. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, π, of a success and probability, 1-π, of a failure remain constant.

Other examples of binomial distributions:

- Counting the number of 2's that are rolled when you roll a die six times (trial = rolling a dice, success = rolling a 2, $n = 6$, $\pi = 1/6 = 0.1667$)
- Counting the number of times a jack is pulled out of deck of cards (with replacement) when you pull a card fifteen times (trial = pulling a card, success = pulling a jack, $n = 15$, $\pi = 4/52 = 0.0769$)
- Counting the number of times that you win a prize in Tim Hortons Roll up the Rim to Win contest out of four cups (trial = checking cup for win, success = winning a prize, $n = 4$, $\pi = 1/6 = 0.1667$ – assuming no special rules (e.g. anniversary rules that changed the odds of winning)

Examples of situations that are not binomial include:

- Counting the number of times a jack is pulled out of deck of cards (without replacement) when you pull a card fifteen times. The fifth criterion is not met because the events are now dependent.
- Counting the number times each number (1 to 6) is rolled when you roll a die fifty times. The fourth criterion is not met because there are six possible outcomes instead of two.
- Counting the number of times that you win a prize in Tim Horton's Roll up the Rim to Win contest out of how many cups you buy during the contest. Unless you know exactly how many you'll buy during the contest, this would not meet the third criterion of having a fixed number of trials.

Note: The Roll up the Rim example might not be binomial as it may fail the fifth criterion. At the beginning of the contest, the odds of winning are determined by counting how many prizes there are out of the total number of

cups printed. As the contest goes on, the probability of winning may change depending on how many people have already won. At the beginning of the contest, this is also true but there are so many cups that it doesn't really matter (think back to the sampling with replacement vs. without replacement in Chapter 1). Thus, this contest is only binomial at the beginning of the contest.

Notation

Suppose we are working on a probability question and there are multiple probabilities that need to be found. Then it gets time consuming to write out, for example, "the probability that three rolls of a die will result in at least one 2" or some variation over and over again. Instead we will use notation to reduce the work. We can write the previous statement more quickly as $P(X \geq 1)$. The $P(\ )$ means the "probability of". X is the random variable being studied (in this case the number of times 2 has been rolled out of 3 rolls). "$X \geq 1$" means we are looking at the number of times a 2 is rolled at least once.

It is important to define X. Otherwise, $P(2<X\leq 5)$ could refer to any random variable and the person reading the notation won't know what it means.

Mean and standard deviation of the binomial distribution

Just like a set of data, a binomial distribution has a mean and a standard deviation. For the binomial distribution, these are given by the formulas:

$\mu = n\pi$

$\sigma = \sqrt{n\pi(1-\pi)}$

Going back to the Tim Hortons example, we had n=4 and π= 0.1667. Thus $\mu = 4 \times 0.1667$= 0.6667 and $\sigma = \sqrt{4 \times 0.1667 \times (1-0.1667)}$ = 0.745. This means that if we buy four random cups of Tim Hortons coffee during the Roll Up the Rim content, we will typically win 0.67 times, give or take 0.75. Thus, when buying four cups of coffee, we will typically win between -0.08 and 1.42 times. Since we can't win negative times, we will round the lower bound to 0. Therefore, when buying four cups of coffee, we will typically win between 0 and 1.42 times.

Exercise:

Problem:

A market research study shows that 30% of all passengers on Canadian Airlines are business travelers. A random sample of 20 passengers is taken.

1. Explain why the above situation satisfies the criteria of a binomial distribution. If there are any issues with why this situation may not meet all of the criteria, discuss them. Define n, X and π.
2. For the random sample, determine the probability that:

a. Exactly seven of the passengers are business travelers.
b. From ten to fourteen (inclusive) are business travelers.
c. At least eleven of these passengers are business travelers.
d. Five of these passengers are NOT travelling on business.

3. What is the typical range of business passengers in a random sample of 20?

Solution:

1. The situation is a binomial distribution because:

- It represents a random variable as the sample is randomly selected.
- It is a discrete variable as we are counting the number of business travellers. *X* is the number of business travellers in the sample.
- There is a fixed number of trials ($n = 20$)
- There are only two options: the passenger is a business traveller (success) or they are not a business traveller (failure).
- In a random sample whether one passenger is a business traveller does not affect the probability of another passenger being a business traveller. Therefore, the probability of success remains constant: $\pi = 30\% = 0.3$

2. Use a computer program to come up with the following output.

X	*P(X)*	$\mathrm{P}(\mathrm{X} \leq \mathrm{x})$
0	0.00080	0.00080
1	0.00684	0.00764
2	0.02785	0.03548
3	0.07160	0.10709
4	0.13042	0.23751
5	0.17886	0.41637
6	0.19164	0.60801
7	0.16426	0.77227
8	0.11440	0.88667
9	0.06537	0.95204
10	0.03082	0.98286
11	0.01201	0.99486
12	0.00386	0.99872
13	0.00102	0.99974
14	0.00022	0.99996
15	0.00004	0.99999
16	0.00001	1.00000
17	0.00000	1.00000

18	0.00000	1.00000
19	0.00000	1.00000
20	0.00000	1.00000

a. P(X=7) = 0.16426
b. P(10 ≤ X ≤ 14) = 0.04792 (highlight the values in the column $P(X)$ for X from 10 to 14, then look at the Sum in the lower right)
c. P(X ≥ 11) = 0.01714 (highlight the values in the column $P(X)$ for X from 11 and higher, then look at the Sum in the lower right)
d. This changes π to 0.7, then re-run the computer program. Look at when X is 5. $P(X=5) = 0.00004$

3. The mean is the same as the expected value, which is 6.0 and the standard deviation is 2.049. This gives us a typical range of 3.951 and 8.049 for the typical number of business passengers in a random sample of 20 passengers.

Evaluating evidence using the binomial distribution

A company looked at its hiring practices. In particular, they found that their hiring practices appears to favour men over women. Based on past data, they have found that regardless of the number of applications by women, seventy-five percent of hires are men. Due to this issue, they decide to implement program. In this program, the name and any identifying features that may indicate the gender of an applicant are removed. For example, if the application says, “She executed a marketing campaign that increased revenue by 30%”, this would be changed to “They executed a marketing campaign that increased revenue by 30%.” The names on the applications were changed an alpha-numeric identification (like AB-101). The company *claims* that the program has worked, but they want to check the claim.

How will the company determine if the program has worked? One way to do this would be using statistics.

Now suppose that after a recent round of hiring, the proportion of men hired was 70%. Would this be enough evidence that the program is working? 70% is definitely lower than 75%, but we know that there is variability in sampling. This means that, prior to the program being implemented, around 75% of hires are men, there may be some rounds of hiring where 70% of hires were men and some that were 80%. It won’t be 75% each time. Instead we expect it to be close to 75%. Therefore, if the program has caused the hiring practices to change, would a recent round of hiring that results in the proportion of men hires being 70% be enough evidence of change? What about 60%? What is the line between normal variability from 75% and abnormal variability? Statistics helps us figure that out and that is how we evaluate evidence using statistics.

Let’s say in a recent round of hires there were 30 new hires and 20 of these hires were men.

Skepticism

Any time we are trying to evaluate evidence, we always start from a position of skepticism. That is, we don’t want to assume what we are trying to show (i.e the claim). If we do that, we may bias the investigation. To illustrate, if you assume that your significant other is cheating on you, then this will colour all of the evidence you find (why did they show up five minutes late from work? They must be cheating!). A well-known real-world example of this position is the assumption in court that a defendant is innocent until proven guilty. That is, criminal court cases start with the assumption of innocence.

In general, the position of skepticism is that nothing has changed, the program didn’t work, the experiment didn’t work, the effect being studied isn’t happening, etc.

In our example, we will assume that the program that the company implemented did not work. That is, we are assuming that the proportion of hires that are men is still 75%. Another way of writing this is $\pi = 75\%$ (i.e. the population proportion).

Evidence

In a court case, evidence would be witness testimony, forensics evidence, expert testimony, etc.

In statistics, evidence is sample data. The evidence has been collected to evaluate the claim. In this case, the evidence has been evaluated to determine if the program is working.

In our example, the sample data is the 20 men hires out of 30. This gives us a sample proportion of 20/30 = 66.67%. The symbol for sample proportion is $\hat{p}$ (said “p hat” - the symbol above the p is supposed to be "^", but the online textbook program does not properly show it).

Evaluating evidence in statistics

To evaluate the evidence, we want to determine the probability of observing the evidence (or even better evidence against the assumption) assuming the assumption is true. Once we determine this probability, we need to determine if the event is unlikely or not unlikely. That is, we want to determine if it unlikely we could have observed the evidence, if the assumption is true. Or is it not unlikely that we observed the evidence, if the assumption is true. If it is unlikely to have observed the evidence, then most likely there is something wrong with the assumption and the claim is likely true. If it is not unlikely to have observed the evidence, then we can’t actually conclude that there is something wrong with the assumption and we cannot conclude that the claim is true.

To go back to the court case example, if you are a juror, you have to evaluate how unlikely or not unlikely it is that the defendant would have had a heated argument with the victim, and was found covered in blood and holding the murder weapon at the scene, if the defendant was innocent. If you think that it is unlikely that all of the pieces of evidence could have happened if the defendant is innocent, then you would find the defendant guilty. That is, the evidence calls into question the assumption. If you think that this it is not unlikely that all of these pieces of evidence could have happened if the defendant is innocent, then you would find the defendant not guilty. Notice that we don’t conclude that the defendant is innocent. That is, we can’t say that they are innocent; we can only say that they are not guilty.

Note: When evaluating evidence, we are trying to evaluate the claim (i.e. not the position of skepticism). Therefore, the evidence has been collected about claim. No evidence has been collected about the assumption. Therefore, our conclusion can only be about the claim and not the assumption.

Therefore, if the probability is small and therefore unlikely, we can say that there is enough evidence to suggest that the assumption is likely false (i.e. guilty).

If the probability is not small and therefore not unlikely, we can say that there is not enough evidence to suggest that the assumption is false (i.e. not guilty).

Note: In statistics, if the probability of an event happening is less than 1%, we say that the event is unlikely to happen. If the probability is greater than 10%, we say the event is not unlikely to happen. If the probability is between 1% and 10%, then it is up to the researcher to determine whether they believe that the event is unlikely or

not unlikely. Usually, the researcher decides on the threshold between unlikely and not unlikely before performing the experiment or study.

In our example, to evaluate the evidence, we want to work out what is the probability this company would have hired 20 men out 30 (or even better evidence against the assumption) if the proportion of men hires is still 75%. That is, we want to find $\mathrm{P}(\mathrm{X} \leq 20)$, given π = 75%). Notice that this is a conditional probability and the condition is the assumption.

What does "or even better evidence against the assumption" mean? It means that we don't just find the probability of exactly 20 out of 30 men hires. We find the probability of at most 20 out of 30 men hires because if the company hired 19, 18, 17, … men then that would be even better evidence that 75% is no longer correct (as the sample proportion is getting more and more different from the assumed population proportion).

Why do we look at "or even better evidence against the assumption"? Often the probability of exactly one event happening is quite small. For example, the probability of getting exactly 10 heads out of 20 coin tosses is 17.62%, even though that is the most likely event to occur. Therefore, if we only looked at the probability of exactly one event happening (i.e. P(X=20)) rather than $\mathrm{P}(\mathrm{X} \leq 20)$, we may come to the false impression that an event is unlikely, when it could actually be explained by normal sampling variability.

Finding the probability

To find the probability, we need to find an appropriate distribution that models the situation. In later chapters, we will look at other models. Right now, the model we are going to use is the binomial distribution. For us to use this model, we have to ensure that the situation is meeting all of the conditions of the binomial distribution.

1. *The variable being studied is random*: This is not necessarily the case here as the applicants are not random and the hiring process is not random. If we randomly selected 30 hires from a greater number of hires, then it would be.
2. *The outcomes of the variable are being counted*: We are counting the number of men hired.
3. *There are a fixed number of trials*: We are looking at 30 hires (n = 30)
4. *There are only two possible outcomes*: Either the hire is a man or the hire is not a man.
5. *The n trials are independent and the probability of success and probability of failure remain constant*: This is true because we are assuming that the probability of hiring a man remains constant at 75%.

Though the first condition is not met, we can still use the binomial distribution to model the situation. That the model is not perfectly met would be a limitation of the study. That means that we would want to put a caveat at the end of our conclusion to state that this might reduce the accuracy of our results.

Note: If the conditions of randomness and independence may not be fully met, then we can still utilize the binomial distribution. But we do have to be wary of the results. The other three conditions do need to be met to use the binomial distribution.

Now that we have the model, we can find the probability. In A computer program, we will use the binomial distribution with n = 30 and π (or probability of occurrence) = 0.75. Then we will find $\mathrm{P}(\mathrm{X} \leq 20)$.

From the computer program, we get $\mathrm{P}(\mathrm{X} \leq 20)$ = 0.19659 = 19.659. Again, this probability is found under the assumption that the program has not worked (i.e. π=75%)

Evaluating the probability

The probability that we would have observed at most 20 hires that were men out of 30, under the assumption that the program did not work, is 19.659%. Therefore, it is *not unlikely* that we could have observed this evidence as the probability is greater than 10%. This means that having 20 out of 30 hires being men falls within the normal sampling variability for this data.

Based on the evidence collected, there is not sufficient evidence to suggest that the program worked. Notice we don't conclude that the program is not working.

Note: In statistics, we never use the words “prove” or “true” when making a conclusion. All of our conclusions are based off of sample data that we are using to make a conclusion about the population. Therefore, there is always the chance of error.

Example

Olivier has spent five years honing his archery skills in various seedy locals around the world. Now he has returned to his city of birth to use these skills to take out criminals. One night while drinking vodka with his friends, he boasts that he can shoot an arrow into the bullseye, blindfolded at a distance of 50m 90% of the time.

“I don’t believe you!” Jack, Olivier’s best friend, slurred.

“I swear! I’ve really honed my skills.” Olivier countered.

“But remember last week when we were in that darkened factory, you missed two of your shots!” Thelma, Olivier’s sister, countered.

“No. I meant to miss them.”

Jack thought for a moment. “I think you are exaggerating and I’m going to test you.”

“You’re on!” Olivier sneered arrogantly.

To test that Olivier was exaggerating about his marksmanship, Jack set up a bunch of targets and, randomly had Olivier attempt the shot. Olivier hit the bullseye (blindfolded at a distance of 50m) 39 out of 50 times.

a. If Olivier’s is not exaggerating, how many times out of 50 do we typically expect him to hit the bullseye? Write your answer as a range that takes into account variation.

Answer: We would expect Olivier to hit the bullseye 45 times give or take 2.121 times. This means a typical range is 42.88 to 47.12 bullseyes out of 50.

b. Based on your answer in a), is 39 out of 50 times potentially abnormal? Explain.

Answer: Since 39 is outside of the range, it would be deemed atypical, but that does not necessarily mean that it is abnormal.

c. What assumption do we need to make before determining whether the 39 out of 50 provides evidence for or against Olivier exaggerating?

Answer: Since Jack wants to show that Olivier is exaggerating, we want to assume that Olivier is not exaggerating. This means we want to assume $\pi = 90\%$, where π is the proportion of bullseyes that Olivier hits.

d. What model (i.e. distribution) will you use to test the evidence against the assumption? Explain why it is the best model to use. Note: This situation might not completely fit the model, but explain why it is still a

reasonable model to use.

Answer: The distribution satisfies the conditions of the binomial distribution:

- *The variable being studied is random*: Since Jack is randomly having Olivier take the shot, we can say this is a random event.
- *The outcomes of the variable are being counted*: We are counting the number of bullseyes.
- *There are a fixed number of trials*: We are looking at 50 shots with the bow and arrow.
- *There are only two possible outcomes*: Either the shot is a bullseye or it is not.
- *The n trials are independent and the probability of success and probability of failure remain constant*: This is true because we are assuming that the probability of hitting the bullseye remains constant at 90%.

e. What probability do you need to find to evaluate the evidence against the assumption?

Answer: We need to find the probability that Olivier hits at most 39 out of 50 bullseyes, assuming his accuracy is 90%. NOTE: We look at "at most 39" because having less bullseyes is even better evidence that Olivier is exaggerating (i.e. better evidence against the assumption).

f. Find that probability.

Answer: $\mathrm{P}(\mathrm{X} \leq 39$ given π=90%) =0.00935=0.94%, (from computer program with n = 50, π (or probability of occurrence) = 90%).

g. In the context of the problem, interpret the probability.

Answer: The probability that Olivier hit at most 39 out of 50 bullseyes, under the assumption that he wasn't exaggerating about his accuracy is 0.94%.

h. Does the probability provide evidence to support whether Olivier is exaggerating or not? Explain.

Answer: Since the probability that we observed our sample data is less than 1%, then it is unlikely that that Olivier is not exaggerating (i.e. that his accuracy is 90%). Therefore, it is likely that Olivier is exaggerating and cannot hit the bullseye 90% of the time blindfolded from 50m.

Exercise:

Problem:

As stated in a previous question, the chance of an CRA audit for a tax return with over $25,000 in income is about 2% per year. An employee at I&S Square, a company that helps individuals do their yearly tax returns and helps if there is an audit, has noticed that people in Seba Beach, Alberta appear to have a greater chance of an audit than the rest of Canadians. Out of a random sample of 45 residents, four of them have been audited.

a. If the residents of Seba Beach are being audited fairly, how many residents out of 45 do we typically expect to get audited in a year? Write your answer as a range that takes into account variation.
b. Based on your answer in a), is 4 out of 45 audits potentially abnormal? Explain.
c. What assumption do we need to make before determining whether the 4 out of 45 audits is unfair?
d. What model (i.e. distribution) will you use to test the assumption? Explain why it is the best model to use. Note: This situation might not completely fit the model, but explain why it is still a reasonable model to use.
e. What probability do you need to find to evaluate the assumption?
f. Find that probability.
g. In the context of the problem, interpret the probability.
h. Does the probability provide evidence to support or refute whether residents of Seba Beach are being unfairly audited? Explain.

Solution:

a. We would expect 0.9 give or take 0.939 to be audited. This means a typical range is 0 to 1.8 residents to be audited out of 45.
b. Since 4 is outside of the range, it would be deemed atypical, but that does not necessarily mean that it is abnormal.
c. Since the employee at I&S Square wants to show that something strange is happening in Seba Beach, they would want to assume that nothing strange is happening. That is, the rate of audits is the same in Seba Beach as anywhere in Canada. This means we want to assume $\pi = 2\%$, where π is the proportion of people who are audited.
d. The binomial distribution:

- *The variable being studied is random*: We are looking at a random sample.
- *The outcomes of the variable are being counted*: We are counting the number of audits.
- *There are a fixed number of trials*: We are looking at 45 residents.
- *There are only two possible outcomes*: Either the resident is audited or they are not.
- *The n trials are independent and the probability of success and probability of failure remain constant*: This is true because we are assuming that the probability of being audited remains constant at 2%.

e. We need to find the probability that at least 4 out of 45 residents are audited, assuming an audit rate of 2%.
f. $P(X \geq 4 \text{ given } \pi = 2\%) = 0.01242 = 1.24\%$ (from computer program with n = 45, π (or probability of occurrence) = 2%).
g. The probability that at least 4 out of 45 Seba Beach residents are audited, under the assumption that the audit rate is 2%, is 1.24%.
h. Since the probability that we observed our sample data is between 1% and 10%, then we have to determine if the probability is unlikely or not unlikely. Since it is closer to 1% than 10%, we can say that the sample data is unlikely to have occurred under the assumption. Therefore, the evidence suggests that there is something wrong with the assumption. That is, there is evidence that the residents of Seba Beach are being audited at a higher rate than the rest of Canada.

Exercise:

Problem:

Executives at Bull, a Canadian own cell phone company, are not very happy with their current customer satisfaction surveys. Using a Likert scale, they surveyed a very large sample of clients who phoned Bull and spoke to a customer service representative. They have determined that only 60% of customers rate their overall satisfaction with the service they received at 4 or higher. That is, they either strongly agree or agreed with the statement, “I am happy with the overall customer service I received during my most recent call to Bull.”

They feel that this is too low as 40% of customers were not happy with their service. To address these issues, they’ve brought in a consultant who has suggested that customers are happier with their service if they feel they’ve built a rapport with the customer service representative. Thus, Bull has decided to train their customer service representatives to start each call with a short conversation. As customers are from across Canada and it would be bad if the conversations were generic, to help their customer service representatives build rapport, a short notice shows up on their screens before they take the call that contains suggested conversation topic for the area the person is calling from. For example, it might include information about weather in the local area and how the local sporting team has done in their most recent game.

After the customer service representatives have been trained in how to make small talk to build rapport, a random sample of sixty customers who called Bull and spoke to a customer service representative is taken. The participants are asked the same question about their overall satisfaction with their customer service phone call as stated above. The results of the survey are listed below:

1	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	3	
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	

Does the recent sample provide sufficient evidence to suggest that the proportion of customers who are happy with their overall service when they call Bull has increased from 60%? Explain your answer in detail.

Solution:

To be skeptical, we want to assume that the program has not worked (i.e. π stayed at 60%). The evidence is 41 out of 60 customers gave a rating of 4 or 5. Perfect evidence that the program worked would be 60 out of 60 happy customers in every single sample. The probability that we would observe at least 41 out of 60 customers who gave a score of four or five regarding their overall satisfaction, assuming that the new program has not worked, is 11.70% (from a computer program with n=60, π=0.60). Therefore, it is not unlikely that we observed the evidence that we did, under the assumption the program did not work. This means that we cannot conclude that the program worked.

Note: What we can conclude when the probability is "not unlikely": If the probability is greater than 10%, then it means that it is not unlikely that we observed this evidence under the assumption. We can NOT conclude that the assumption is likely true as the evidence was collected to evaluation the claim (not the assumption). Instead, we can only conclude that there is not enough evidence to say that the claim is true. When the probability is "not unlikely", we have really learned very little about the claim.

Chapter Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. The variable being studied is random.
2. The outcomes of the variable are being counted.
3. There are a fixed number of trials. The letter n denotes the number of trials.
4. There are only two possible outcomes, called "success" and "failure," for each trialπ denotes the probability of a success on one trial, and 1-π denotes the probability of a failure on one trial.
5. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, π, of a success and probability, 1-π, of a failure remain constant.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X = the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = n\pi$, and the standard deviation is given by the formula $\sigma = \sqrt{n\pi(1-\pi)}$.

To evaluate evidence, we must first begin from a position of skepticism (i.e. assume the opposite of what we want to show). Then we must find a probability which is the distance from the actual evidence to perfect evidence against the assumption. We can then evaluate the probability by determining whether it is less than 1% (which means it is unlikely the evidence occurred under the assumption) or if it is greater than 10% (which means it is not unlikely the evidence occurred under the assumption). If the probability is deemed unlikely, then we reject the assumption, which means there is enough evidence to support what we originally wanted to show (the claim). If the probability is deemed not unlikely, then we do not reject the assumption, which means there is not enough evidence to support what we originally wanted to show (the claim). In the latter situation, we cannot make any conclusions about the assumption as the evidence was collected only for the claim.

Practice

The first few exercises provided are from the textbook Business Statistics -- BSTA 200 -- Humber College -- Version 2016RevA -- DRAFT 2016-04-04 by Alexander Holmes, Lyryx Learning: http://cnx.org/contents/f3aefa9e-58d2-41ea-969f-04dc2cb04c82@5.20

Use the following information to answer the next seven exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

Exercise:

Problem: In words, define the random variable *X*.

Solution:

X = the number that reply “yes”

Exercise:

Problem: What values does the random variable *X* take on?

Solution:

0, 1, 2, 3, 4, 5, 6, 7, 8

Exercise:

Problem:

Construct the probability distribution function (PDF). That is, fill in the table below. In the left column put in the possible values for X. In the right column, put in the probability for exactly X, i.e. $\mathrm{P}(\mathrm{X} = \mathrm{x})$

x	$\mathrm{P}(\mathrm{X} = \mathrm{x})$

Solution:

x	$P(X = x)$
0	0.00005
1	0.0009
2	0.0080
3	0.0395
4	0.1227
5	0.2439
6	0.3030
7	0.2151
8	0.0668

Exercise:

Problem: On average (μ), how many would you expect to answer yes?

Solution:

5.7

Exercise:

Problem: What is the standard deviation (σ)?

Solution:

1.2795

Exercise:

Problem: What is the probability that at most five of the freshmen reply “yes”?

Solution:

0.4151

Exercise:

Problem: What is the probability that at least two of the freshmen reply “yes”?

Solution:

0.9990

Exercise:

Problem:

A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

a. In words, define the random variable X.
b. List the values that X may take on.
c. How many of the 12 students do we expect to attend the festivities?
d. Find the probability that at most four students will attend.
e. Find the probability that more than two students will attend.

Solution:

a. X = the number of students who will attend Tet.
b. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
c. 2.16
d. 0.9511
e. 0.3702

Use the following information to answer the next three multiple choice questions: The probability that the Calgary Flames will win any given game is 0.4617 based on a 45-year win history of 1,616 wins out of 3,500 games played (as of Sept. 2017). An upcoming monthly schedule contains 12 games.

Exercise:

Problem: The expected number of wins for that upcoming month is:

a. 1.67
b. 12
c. $\frac{1616}{3500}$
d. 5.54

Solution:

d. 5.54

Let X = the number of games won in that upcoming month.

Exercise:

Problem: What is the probability that the Calgary Flames win exactly six games in that upcoming month?

a. 0.2178
b. 0.4167
c. 0.7664
d. 0.7116

Solution:

a

Exercise:

Problem: What is the probability that the Calgary Flames win at least five games in that upcoming month

a. 0.2176
b. 0.2762
c. 0.7238
d. 0.5062

Solution:

c

Exercise:

Problem:

The chance of an Canadian Revenue Agency audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

a. In words, define the random variable X.
b. List the values that X may take on.
c. How many audits are expected in a 20-year period?
d. Find the probability that a person is not audited at all.
e. Find the probability that a person is audited more than twice.

Solution:

a. X = the number of audits in a 20-year period
b. 0, 1, 2, …, 20
c. 0.4
d. 0.6676
e. 0.0071

Exercise:

Problem:

According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X = the number of people who have access to electricity.

a. Calculate the mean and standard deviation of X.
b. Find the probability that 15 people in the sample have access to electricity.
c. Find the probability that at most ten people in the sample have access to electricity.
d. Find the probability that more than 25 people in the sample have access to electricity.

Solution:

a. Mean = np = 150(0.09) = 13.5; Standard Deviation = $\sqrt{npq}$ = $\sqrt{150(0.09)(0.91)} \approx 3.5050$
b. $P(x = 15) = 0.0988$
c. $P(x \leq 10) = 0.1987$
d. $P(x > 25) = 0.0009$

Exercise:

Problem: Jenna and Megan looked at the new packaging.

“I guess it looks ok.” Megan hedged.

“The design team says that this new packaging really sells the time-saving nature of the kit.”

“But it’s kinda off-putting.” They continued to stare at the new packaging. Jenna and Megan had developed a make-up kit called ‘5 minute make-up’, which was aimed at women on the go who wanted to put ‘their face on’ but a lot quicker than they usually did. Their target market was new moms, moms with full-time jobs, full-time students working full-time, …. In other words, anyone who didn’t have 30 minutes every morning to do their make-up. Their little start-up was doing well. They’d arranged for their product to be produced, had made how-to videos on YouTube, and were starting to get their products put into stores. Their dream placement was in Sephora.

Now that they were more established, they had decided to hire a marketing expert who could help take them to the next level. The first thing that Leticia suggested was to change the packaging. She argued that their old packaging didn’t convey the premise of the product clearly enough. With the help of a design team, Leticia had come up with a packaging that showed a harried woman with hair everywhere and bags under her eyes looking overwhelmed. But when you flip the package over, the woman was now perfectly put together – ‘only five-minute make-up can save you from being a hot mess’.

Jenna finally broke the silence. “I don't know if I would even pick up this package. It just looks so depressing. But what do we do?”

“Leticia wants to put the product in this new packaging in five stores that carry our products. Based off of previous sales numbers, we know that the stores sell 68% of the product we give them in a two-week period.”

“How does that help us? Do we just watch our sales plummet?” Jenna was sounding exasperated.

“I’m getting to that.” Megan soothed. “Leticia is convinced this packaging will increase sales. But what if we can show her that it doesn’t? Let’s put this packaging into the five stores and then see how many kits were actually sold. I bet that we can show her that the sales went down.”

“I don’t see how that is useful. We still have to pay her stupid fee.”

“You should read her contract more closely. She only gets paid if she can show that sales increased. If they don’t, then not only does she not get paid but she also has to pay for any contractors (i.e. the design team).”

Jenna perked up visibly at this.

Over the next two weeks, five stores carried the new packaging. Megan and Jenna provided each store with 100 kits. At the end of the two weeks, 306 of the kits were sold.

1. What is the observation unit? What is the variable? Categorize it.
2. What do Jenna and Megan want to show?
3. What assumption do Jenna and Megan need to make in order to investigate your answer in question 2? Write your answer both as a sentence and as a probability.
4. What is the evidence that Jenna and Megan have found?
5. Describe the process that Jenna and Megan will go through to evaluate this evidence. Your description should include (but is not limited to) what probability they will find and what they will do with that probability once they’ve found it. Don’t actually do the process (that comes later). Just describe what they will do.
6. Jenna and Megan believe that the binomial distribution will be the best model to find the required probability. Does this situation meet the criteria for a binomial? Examine each criterion and comment on whether it is satisfied here or not.
7. Regardless of your answer above, use a binomial distribution to model this situation. Find the appropriate probability to evaluate the evidence using MegaStat.

8. In sentence form, explain what the probability you have found means in the context of the question. Do not make a conclusion yet. Instead explain what it is a probability of.
9. Now make a conclusion. In particular, answer this question: Is their enough evidence to suggest that Leticia's new packaging has reduced sales? Justify your answer.

Solution:

1. Observational unit: Five-minute make-up kit; Variable: Did it sell or not; Categorize: Categorical
2. They want to show that the new packaging will decrease sales.
3. They need to assume the opposite of what they want to show. Therefore, they need to assume that the new packaging does not decrease sales. Therefore, the proportion of kits sold stays the same at 68%.
4. They have found that out of 500 kits supplied, 306 of them have been sold.
5. They first need to start with an assumption (= 68%). Then they need to come up with a model based on this assumption. Once they have the model, they will use it to find the probability that the stores sold at most 306 out of 500 kits, assuming that the new packaging has not decreased sales (i.e. stayed at 68%). Once they have the probability, they need to determine whether the event is likely or unlikely. An event is unlikely if the probability is less than 1%. An event is likely if the probability is more than 10%. If the event is unlikely, then it means that it is unlikely we observed the evidence under the assumption. Since we know the evidence actually happened, that makes us question the assumption. Thus, it is unlikely the assumption is true based off of the evidence. If the event is likely to happen, then the assumption is likely to be true based off of the evidence.

6.
   - Is the data randomly collected? Most likely not. The 500 kits that we are looking at were not randomly selected.
   - Is the data discrete (countable)? As we are counting the number of kits that are sold the data is discrete.
   - Are the events independent? This may be a fair assumption for this study. Most likely the sale of one kit is not dependent on whether another kit is sold. Though if two friends buy the kits together or someone buys a bunch as presents, this is not the case, but in general it is more likely independent than dependent.
   - Are there a fixed number of trials? In this case, the number of trials would be the 500 kits with the new packaging.
   - Are there two possible outcomes? Either a kit is sold or it is not.

7. $P(X \leq 36)$ = 0.00077 = 0.077%
8. The probability that we observed at most 306 out of 500 kits sold (1), assuming the rate of sales is 68%, is 0.077%
9. Since the probability is less than 1%, then it is very unlikely that we would have observed this evidence under the assumption. Since we actually observed the evidence but assumed that the rate was 68%, what we have assumed is called into question. Therefore, it is unlikely that the assumption is true . Therefore, it is likely that the new packaging has resulted in a decrease in sales .

Exercise:

Problem:

Striking Donkey Coffee recently sold an 80% stake in their company to Baravalle, an Italian coffee conglomerate. Striking Donkey's logo is simplistic. Baravalle wants to maintain brand recognition, but also wants to put their stamp on the company. In particular, Baravalle is known for its modern and stylish advertisements.

Designers and marketers at Baravalle have worked tirelessly for the last month to come up with two revised Striking Donkey logos (not included, because it is top secret). They are referred to as Logo 1 and Logo 2.

Now they want to determine whether customers show any preference to either logo. To do this, they asked a random sample of 40 customers who were familiar with the Striking Donkey Brand which logo they prefer.

Participants had to make a choice between the logos.

The results of the study were that 26 out of the 40 participants preferred Logo 2.

The marketers at Baravalle now want to do a statistical analysis to determine whether Logo 2 is preferred significantly more than Logo 1.

1. What assumption do you need to start with when determining whether Logo 2 is preferred significantly more than Logo 1? State your answer both in a sentence and mathematically.
2. Can this situation be modelled by the binomial distribution? Support your answer by showing why or why not this situation satisfies each of the five criteria of the binomial distribution.
3. After previous issues with horrible new logo launches, Baravalle only wants to go forward if there is clear evidence that Logo 2 is preferred. Based on this, what level of significance should they use? Explain your reasoning.
4. Regardless of your answer in b, assume that this situation satisfies the binomial distribution for the remainder of the question. Use a computer program to find that appropriate probability that will allow you to evaluate the evidence.
5. In sentence form, explain what the probability you have found means in the context of the question. Do not make a conclusion yet. Instead explain what it is a probability of.
6. Based on the probability, determine whether Logo 2 is preferred significantly more than Logo 1. Explain your reasoning.

Solution:

Observational unit: People who are familiar with Striking Donkey Coffee; Variable: Whether they prefer Logo 2; Type of variable: Categorical.

1. They need to assume the opposite of what they want to show. This means they need to assume that Logo 2 is NOT preferred significantly more than Logo 1. This would mean they are preferred equally. Therefore, there is a 50% chance that someone will choose Logo 2.

2.
   - The data is collected randomly: Yes. It is a random sample of participants.
   - The outcomes are counted: Yes. They count how many people like Logo 2.
   - There are two possible outcomes: Yes. Either they prefer Logo 2 or they did not.
   - There are a fixed number of trials: Yes. They asked 40 people.
   - The trials are independent of each other. Yes. It is fair to assume that no participant's preference is based on another participants preference.

3. The more unlikely it is that we observed our evidence, the smaller the probability will be. This means, the smaller the probability, the more unlikely it is that the assumption (i.e. that there is no preference between the logos) is true. Since the marketers want clear evidence that there is a preference, they want a smaller probability, which would show it is unlikely that there is a preference between the logos. The level of significance is the threshold between likely and unlikely. Thus, if they want clear evidence, they want to set their threshold "high", meaning they want to make it a small number. Since the level of significance is between 1% and 10%, the lowest level of significance (meaning the highest threshold of evidence) is at 1%.
4. $P(X \geq 26) = 0.04035 = 4.04\%$
5. The probability that we observed at least 26 out of 40 people who preferred Logo 2, assuming that there was no preference between the logos, is 4.04%.
6. Since the probability is greater than 1% (it is 4.04%), it is not unlikely that we observed at least 26 out of 40 people who preferred Logo 2, assuming that there was no preference between the logos. Therefore, we do not reject that there was no preference between the logos. This suggests that Logo 2 is NOT preferred significantly more than Logo 1.

Introduction -- The Normal Distribution -- Mt Royal University -- Version 2016RevA
class="introduction"

If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

Note:
Chapter Objective
By the end of this chapter, the student should be able to:

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

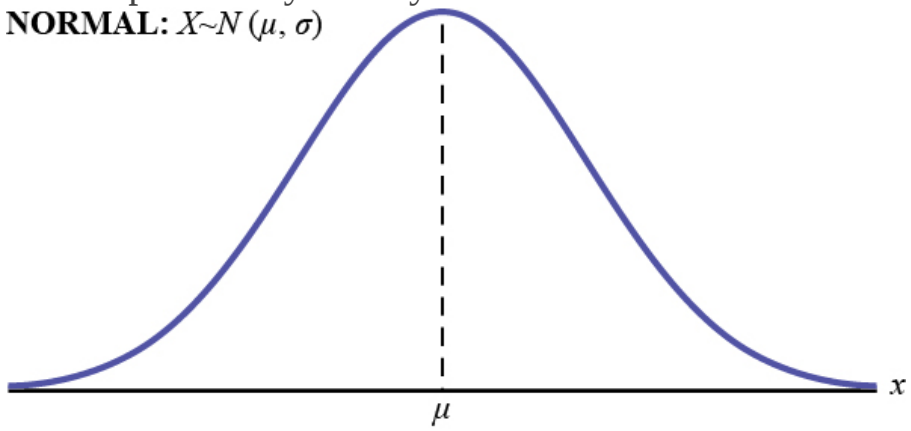
The normal probability density function, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all

disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution.

The normal distribution is extremely important, but it cannot be applied to everything in the real world. Remember here that we are still talking about the distribution of population data. This is a discussion of probability and thus it is the population data that may be normally distributed, and if it is, then this is how we can find probabilities of specific events just as we did for population data that may be binomially distributed or Poisson distributed. This caution is here because in the next chapter we will see that the normal distribution describes something very different from raw data and forms the foundation of inferential statistics.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing the following formula of the normal probability density function:



The curve is symmetrical about a vertical line drawn through the mean, μ. The mean is the same as the median, which is the same as the mode, because the graph is symmetric about μ. As the notation indicates, the

normal distribution depends only on the mean and the standard deviation. Note that this is unlike several probability density functions we have already studied, such as the Poisson, where the mean is equal to μ and the standard deviation simply the square root of the mean, or the binomial, where p is used to determine both the mean and standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter and wider or skinnier and taller depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

Formula Review

$X \sim N(\mu, \sigma)$

μ = the mean σ = the standard deviation

Glossary

Normal Distribution

a continuous random variable (RV) with pdf $f(x)$ =

$$\frac{1}{\sigma\sqrt{2\pi}} \, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV, Z, is called the **standard normal distribution**.

The Standard Normal Distribution-- The Normal Distribution --MRU - C Lemieux

The **standard normal distribution** is a normal distribution of **standardized values called *z*-scores**. **A *z*-score is measured in units of the standard deviation.** For example, if the mean of a normal distribution is five and the standard deviation is two, the value x = 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$

The *z*-score is three.

The mean for the standard normal distribution is zero, and the standard deviation is one. What this does is dramatically simplify the mathematical calculation of probabilities. Take a moment and substitute zero and one in the appropriate places in the above formula and you can see that the equation collapses into one that can be much more easily solved using integral calculus. The transformation $z = \frac{x-\mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value *x* comes from a known normal distribution with known mean μ and known standard deviation σ. The *z*-score tells how many standard deviations a particular x is away from the mean.

Z-Scores

If *X* is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z*-score is:
Equation:

$$z = \frac{x - \mu}{\sigma}$$

The *z*-score tells you how many standard deviations the value *x* is above (to the right of) or below (to the left of) the mean, μ. Values of *x* that are larger than the mean have positive *z*-scores, and values of *x* that are smaller than the mean have negative *z*-scores. If *x* equals the mean, then *x* has a *z*-score of zero.

Example:
Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then:
Equation:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **two standard deviations** (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.
Now suppose $x = 1$. Then: $z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$ (rounded to two decimal places)
This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$.

Example:
Some doctors believe that a person can lose five pounds, on average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$.
Exercise:

Problem:

Suppose a person **gained** three pounds (a negative weight loss). Then $z =$ _____. This z-score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution:
Equation:

$$Z = \frac{x - \mu}{\sigma} = \frac{-3 - 5}{2} = -4$$

z = **–4**. This z-score tells you that x = –3 is **four** standard deviations to the **left** of the mean.

Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

$z = \frac{y-\mu}{\sigma} = \frac{4-2}{1} = 2$ where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of **their** respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Note:
Try It
Exercise:

Problem: Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16,4)$. Suppose Jerome scores ten points in a game. The z–score when $x = 10$ is –1.5. This score tells you that $x = 10$ is _____ standard deviations to the _____(right or left) of the mean_____(What is the mean?).

Solution:

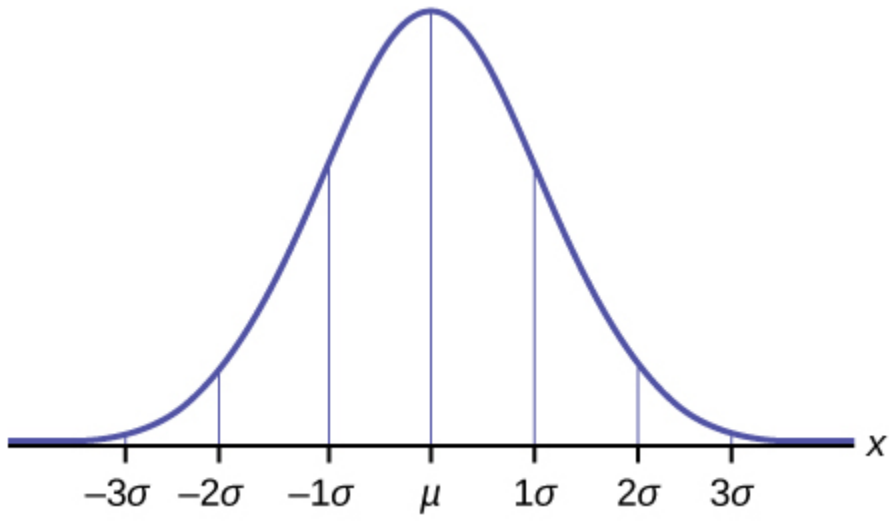
1.5, left, 16

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** says the following:

- About 68.26% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95.44% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.73% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for $+1\sigma$ and -1σ are +1 and –1, respectively.
- The z-scores for $+2\sigma$ and -2σ are +2 and –2, respectively.
- The z-scores for $+3\sigma$ and -3σ are +3 and –3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.



Example:
The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Exercise:

Problem:

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The *z*-score when $x = 168$ cm is $z =$ _____. This *z*-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Solution:
Equation:

$$Z = \frac{x - \mu}{\sigma} = \frac{168 - 170}{6.28} = -0.32$$

a. –0.32, 0.32, left, 170

Exercise:

Problem:

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a *z*-score of $z = 1.27$. What is the male's height? The *z*-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Solution:
Equation:

$$Z = \frac{x - \mu}{\sigma} = \frac{x - 170}{6.28} = 1.27 \rightarrow 1.27*6.28 + 170 = 177.98$$

b. 177.98, 1.27, right

Note:
Try It
Exercise:

Problem:

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Solution:

The z-score for $x_1 = 325$ is $z_1 = –1.14$.

The z-score for $x_2 = 366.21$ is $z_2 = –1.14$.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Example:
Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie between $–1\sigma = (–1)(6) = –6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 – 6 = 44$ and $50 + 6 = 56$ are

within one standard deviation of the mean 50. The *z*-scores are –1 and +1 for 44 and 56, respectively.

- About 95% of the *x* values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values 50 – 12 = 38 and 50 + 12 = 62 are within two standard deviations of the mean 50. The *z*-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the *x* values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values 50 – 18 = 32 and 50 + 18 = 68 are within three standard deviations of the mean 50. The *z*-scores are –3 and +3 for 32 and 68, respectively.

Note:
Try It
Exercise:

Problem:

Suppose *X* has a normal distribution with mean 25 and standard deviation five. Between what values of *x* do 68% of the values lie?

Solution:

between 20 and 30.

Note:
Try It
Exercise:

Problem:

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68% of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.
b. About 95% of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.
c. About 99.7% of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.

Solution:

a. About 68% of the values lie between the values 41 and 63. The *z*-scores are –1 and 1, respectively.
b. About 95% of the values lie between the values 30 and 74. The *z*-scores are –2 and 2, respectively.
c. About 99.7% of the values lie between the values 19 and 85. The *z*-scores are –3 and 3, respectively.

References

"Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).

"The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

"2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).

“Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009.” National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

“List of stadiums by capacity.” Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Chapter Review

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Practice Questions

Exercise:

Problem:

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution:

0.67, right

Exercise:

Problem:

In a normal distribution, $x = –5$ and $z = –3.14$. This tells you that $x = –5$ is ____ standard deviations to the ____ (right or left) of the mean.

Solution:

3.14, left

Exercise:

Problem:

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Solution:

about 68%

Exercise:

Problem:

About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Solution:

about 95.44%

Exercise:

Problem:

About what percent of x values lie between the second and third standard deviations (both sides)?

Solution:

about 4%

Use the following information to answer the next two multiple choice exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Exercise:

Problem: What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

Solution:

b

Exercise:

Problem:

What is the *z*-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

Solution:

c

Exercise:

Problem:

Wesley Crusher was tasked with exploring the **Selcundi Drema** sector. He found a new species of tribbles. In his final report, he stated, "Though tribbles vary in size and dimension, the middle 99.73% of them weigh between 4 and 7.2 kg and follow a normal distribution." Based on this, what is the mean and standard deviation for the weight of tribbles? Choose the best answer.

a. mean = 5.6 kg, standard deviation = 1.07 kg
b. mean = 5.6 kg, standard deviation = 0.53 kg
c. mean = 5.6 kg, standard deviation = 0.8 kg
d. mean = 99.73 kg, standard deviation = 3.2 kg
e. There is not enough information to determine this.

Solution:

b

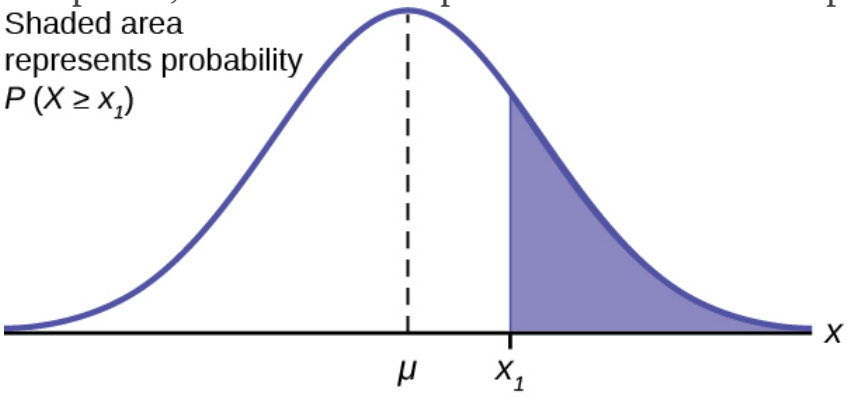
Glossary

Standard Normal Distribution
: a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

z-score
: the linear transformation of the form $z = \frac{x-\mu}{\sigma}$ or written as $z = \frac{|x-\mu|}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0,1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently. A z-score is the number of standard deviations a particular x is away from its mean value.

Using the Normal Distribution-- The Normal Distribution -- MRU - C Lemieux

The shaded area in the following graph indicates the area to the right of x. This area is represented by the probability $P(X > x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X > x)$.

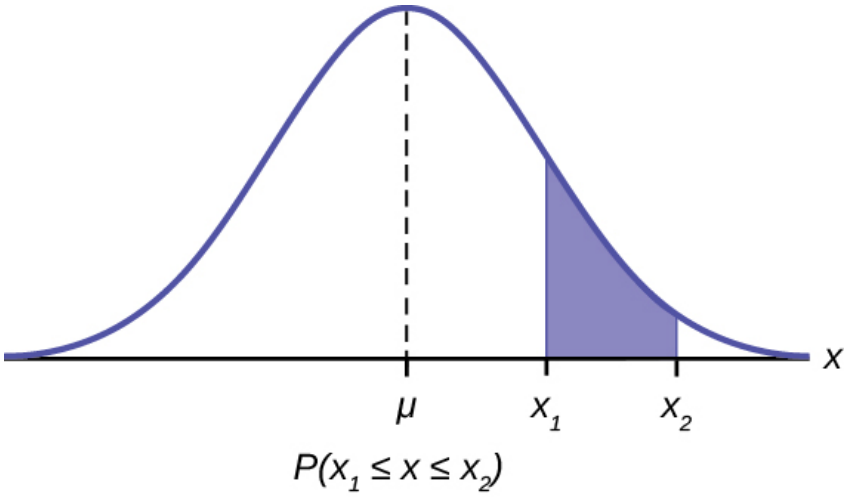


The area to the right is then $P(X > x) = 1 – P(X < x)$. Remember, $P(X < x)$ = **Area to the left** of the vertical line through x. $P(X < x) = 1 – P(X < x)$ = **Area to the right** of the vertical line through x. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

To find the probability for probability curves with a continuous random variable we need to calculate the area under the curve across the values of X we are interested in. For the normal distribution this seems a difficult task given the complexity of the formula. There is, however, a simply way to get what we want.

We start knowing that the area under a probability curve is the probability.



This shows that the area between X_1 and X_2 is the probability as stated in the formula: P ($X_1 \leq x \leq X_2$)

The mathematical tool needed to find the area under a curve is integral calculus. The integral of the normal probability density function between the two points x_1 and x_2 is the area under the curve between these two points and is the probability between these two points.

Doing these integrals is no fun and can be very time consuming. But now, remembering that there are an infinite number of normal distributions out there, we can consider the one with a mean of zero and a standard deviation of 1. This particular normal distribution is given the name Standard Normal Distribution. Putting these values into the formula it reduces to a very simple equation. We can now quite easily calculate all probabilities for any value of x, for this particular normal distribution, that has a mean of zero and a standard deviation of 1. These have been produced and are available here in the text or everywhere on the web. They are presented in various ways. The table in this text is the most common presentation and is set up with probabilities for one-half the distribution beginning with zero, the mean, and moving outward. The shaded area in the graph at the top of the table represents the probability from zero to the specific Z value noted on the horizontal axis, Z.

The only problem is that even with this table, it would be a ridiculous coincidence that our data had a mean of zero and a standard deviation of one. The solution is to convert the distribution we have with its mean and

standard deviation to this new Standard Normal Distribution. The Standard Normal has a random variable called Z.

Using the standard normal table, typically called the normal table, to find the probability of one standard deviation, go to the Z column, reading down to 1.0 and then read at column 0. That number, 0.3413 is the probability from zero to 1 standard deviation. At the top of the table is the shaded area in the distribution which is the probability for one standard deviation. The table has solved our integral calculus problem. But only if our data has a mean of zero and a standard deviation of 1.

However, the essential point here is, the probability for one standard deviation on one normal distribution is the same on every normal distribution. If the population data set has a mean of 10 and a standard deviation of 5 then the probability from 10 to 15, one standard deviation, is the same as from zero to 1, one standard deviation on the standard normal distribution. To compute probabilities, areas, for any normal distribution, we need only to convert the particular normal distribution to the standard normal distribution and look up the answer in the tables. As review, here again is the **standardizing formula**:

Equation:

$$Z = \frac{x - \mu}{\sigma}$$

where Z is the value on the standard normal distribution, X is the value from a normal distribution one wishes to convert to the standard normal, μ and σ are, respectively, the mean and standard deviation of that population. Note that the equation uses μ and σ which denotes population parameters. This is still dealing with probability so we always are dealing with the population, with **known** parameter values and a **known** distribution. It is also important to note that because the normal distribution is symmetrical it does not matter if the z-score is positive or negative when calculating a probability. One standard deviation to the left (negative Z-score) covers the same area as one standard deviation to the right (positive Z-score). This fact is why the Standard Normal tables do not provide areas for the left side of

the distribution. Because of this symmetry, the Z-score formula is sometimes written as:

Equation:

$$Z = \frac{|x - \mu|}{\sigma}$$

Where the vertical lines in the equation means the absolute value of the number.

What the standardizing formula is really doing is computing the number of standard deviations X is from the mean of its own distribution. The standardizing formula and the concept of counting standard deviations from the mean is the secret of all that we will do in this statistics class. The reason this is true is that **all** of statistics boils down to variation, and the counting of standard deviations is a measure of variation.

This formula, in many disguises, will reappear over and over throughout this course.

Example:

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

Exercise:

Problem:

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.

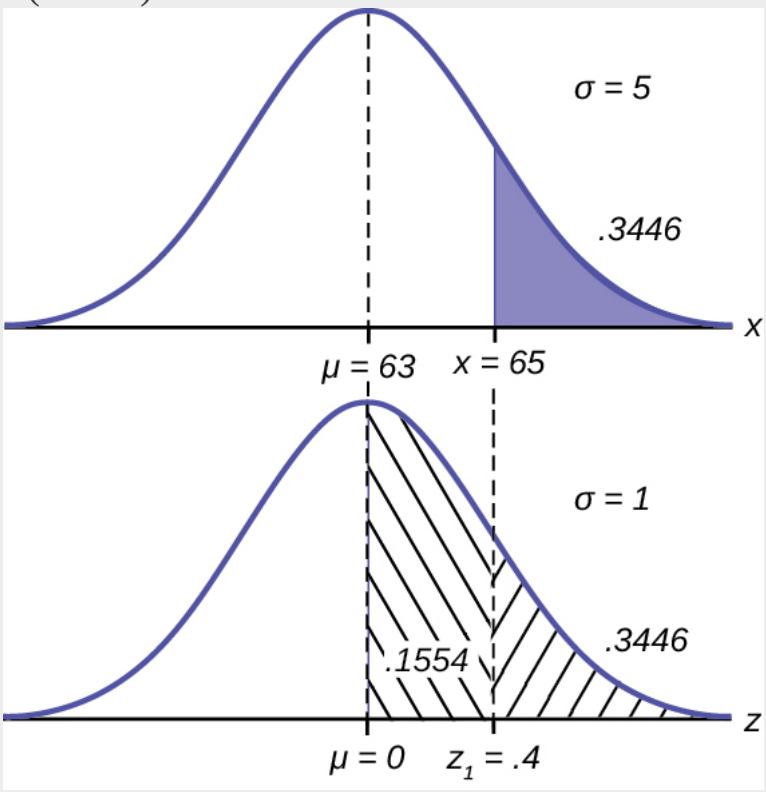
Solution:

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$



Equation:

$$Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{65 - 63}{5} = 0.4$$

$P(x \geq x_1) = P(Z \geq Z_1) = 0.3446$

The probability that any student selected at random scores more than 65 is 0.3446. Here is how we found this answer.

The normal table provides probabilities from zero to the value Z_1. For this problem the question can be written as: $P(X \geq 65) = P(Z \geq Z_1)$,

which is the area in the tail. To find this area the formula would be 0.5 – P(X ≤ 65). One half of the probability is above the mean value because this is a symmetrical distribution. The graph shows how to find the area in the tail by subtracting that portion from the mean, zero, to the Z_1 value. The final answer is: P(X ≥ 63) = P(Z ≥ 0.4) = 0.3446

$z = \frac{65-63}{5} = 0.4$

Area to the left of Z_1 to the mean of zero is 0.1554

$P(x > 65) = P(z > 0.4) = 0.5 - 0.1554 = 0.3446$

Exercise:

Problem:

Solution:

b.

$Z = \frac{x-\mu}{\sigma} = \frac{85-63}{5} = 4.4$ which is larger than the maximum value on the Standard Normal Table. Therefore, the probability that one student scores less than 85 is approximately one or 100%.

A score of 85 is 4.4 standard deviations from the mean of 63 which is beyond the range of the standard normal table. Therefore, the probability that one student scores less than 85 is approximately one (or 100%).

Note:
Try It
Exercise:

Problem:

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

Find the probability that a randomly selected golfer scored less than 65.

Solution:

normalcdf(10^{99},65,68,3) = 0.1587

Example:
A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Exercise:

Problem:

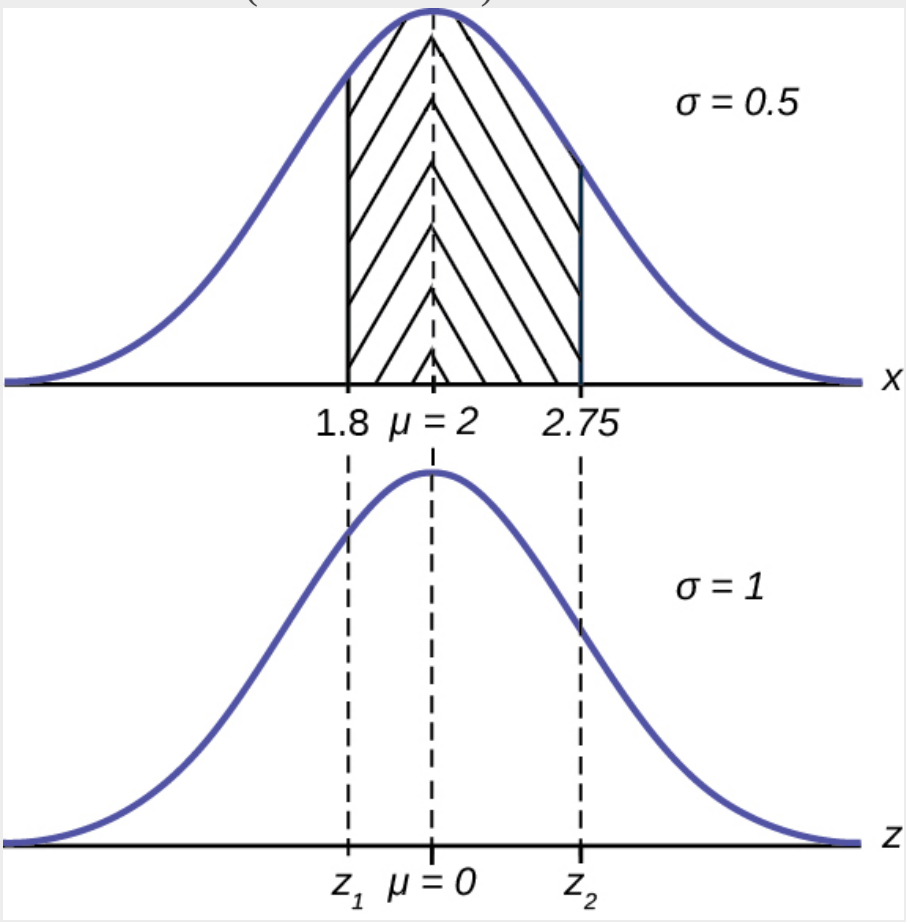
a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Solution:

a. Let X = the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$



$P(1.8 \leq x \leq 2.75) = P(Z_i \leq Z \leq Z_2)$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

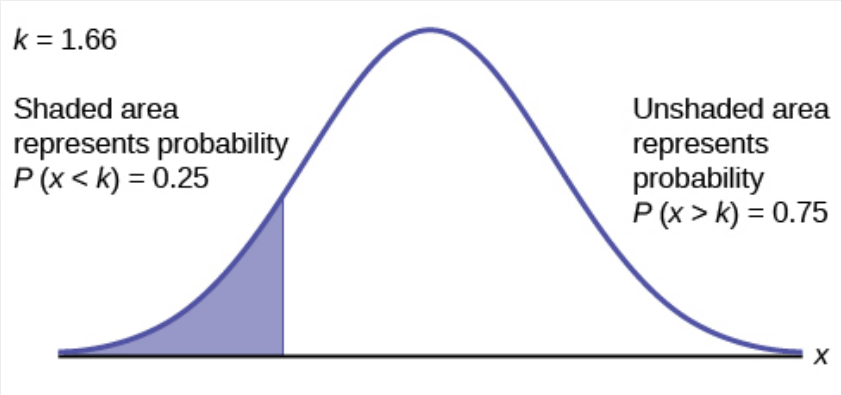
Exercise:

Problem:

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution:

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** *k*, where $P(x < k) = 0.25$.



$f(Z) = 0.5 - 0.25 = 0.25$, *therefore* $Z \approx -0.675$(*or just 0.67 using the table*)$Z = \frac{x-\mu}{\sigma} = \frac{x-2}{0.5} = -0.675$, *therefore* $x = -0.675 * 0.5 + 2 = 1.66$ *hours.*

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Note:
Try It
Exercise:

Problem:

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Solution:

normalcdf(66,70,68,3) = 0.4950

Example:
There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

Exercise:

Problem:

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

Solution:

a. 0.8186

Exercise:

Problem:

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

Solution:

b. 0.8413

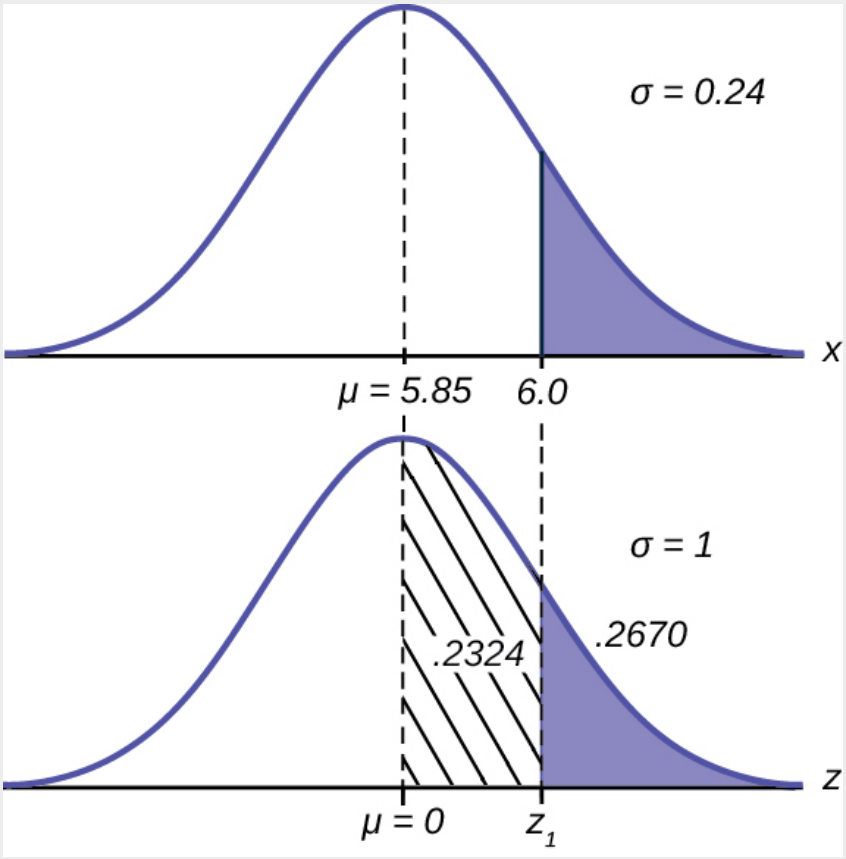
Example:
A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

Exercise:

Problem:

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.

Solution:



Equation:

$$Z_1 = \frac{6 - 5.85}{.24} = .625$$

$P(x \geq 6) = P(z \geq 0.625) = 0.2670$

b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.

Exercise:

Problem:

Solution:

$f\ Z\ = \frac{0.20}{2} = 0.10$, *therefore Z* $\approx \pm 0.25$
$Z = \frac{x-\mu}{\sigma} = \frac{x-5.85}{0.24} = \pm 0.25 \rightarrow \pm 0.25 * 0.24 + 5.85 = \ 5.79, 5.91$

References

“Naegele’s rule.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).

“403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radio-archives/episode/403/nummi (accessed May 14, 2013).

“Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at http://www.winatthelottery.com/public/department40.cfm (accessed May 14, 2013).

“Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).

“Facebook Statistics.” Statistics Brain. Available online at http://www.statisticbrain.com/facebook-statistics/(accessed May 14, 2013).

Practice questions

Exercise:

Problem:

A local bank has determined that the daily balances of the chequing accounts of their customers are normally distributed with a mean of $280 and a standard deviation of $20.

a. What percentage of their customers has daily balances less than $290?
b. What percentage of their customers has daily balances between $250 and $275?
c. What percentage of their customers has daily balances over $260?
d. The Bank is planning a special promotion where it is rewarding its customers whose balances are in the top 15% with a free toaster. What account balance must a customer achieve in order to qualify for a free toaster?
e. 68.26% of balances will be between what amount?
f. What is the interquartile range for the account balances?

Solution:

a. 0.6915
b. 0.3345
c. 0.8413
d. $300.70
e. $260 to $300
f. IQR = 293.5-266.5 = $27

Exercise:

Problem:

The Old Baldy Tire Company is introducing a new steel belted radial tire. Old Baldy's engineering department has estimated that the average life of this tire will be 50,000 km and that the standard deviation of these tires will be 5000 km. It is assumed that the useful life of these tires follows a normal distribution.

a. What is the probability that:

i. These tires will last for longer than 60,000 km?
ii. These tires will last for less than 38,000 km?
iii. These tires will last for between 45,000 and 58,000 km?
iv. These tires will last for between 39,000 and 43,000 km?

b. The Old Baldy Tire Company is considering offering a tire guarantee that each new set of tires will last a certain number of kilometers. If the tires fail to last the specified number of kilometers a new set of tires will be provided to the purchaser for free. The Old Baldy Tire Company wants to ensure that no more than 10% of the tires produced qualify for this guarantee. For how many kilometers should these tires be guaranteed to last?
c. 35% of tires will last less than how many kilometers?

Solution:

a.
i. 0.0228
ii. 0.0082
iii. 0.7865
iv. 0.0669

b. 56407.8 km
c. 48073.4 km

Introduction - Sampling distributions - MRU - C Lemieux
Introduction to section on sampling distributions

Note:
Chapter Objective
By the end of the chapter, the student should be able to:

- Distinguish between a sample, a population and a sampling distribution.
- Know the characteristics of the sampling distribution.
- Recognize sampling distribution problems.
- Apply and interpret the central limit theorem for both means and proportions.

Introduction to Sampling Distributions
Introduces the concept of sampling distributions.

When we take a random sample from a population, we expect that there is going to be some variability (i.e. sampling variability) between the information the sample gives us and the whole population. That is, we might find that the sample mean and the population mean are different. We may also find that if we take multiple random samples of size *n* that the sample mean for each sample is different. The following chapter looks at how we can better understand the sampling variability in statistics.

Before we go on, here is a reminder of a few terms and symbols.

A **parameter** is a descriptive measure of the population (eg. population mean, population standard deviation, population proportion).

A **statistic** is a descriptive measure of the sample (eg. sample mean, sample standard deviation, sample proportion).

Measure	**Population**	**Sample**
Sample size	N	n
Mean	μ_x	$\bar{x}$
Standard deviation	σ_x	s_x
Proportion	π	$\hat{p}$

Table of important symbols

The population mean, population standard deviation, and sample standard deviation have a subscript of x to demonstrate that they are the measure for the variable X. Though this is mostly notational, it does become important later in this chapter.

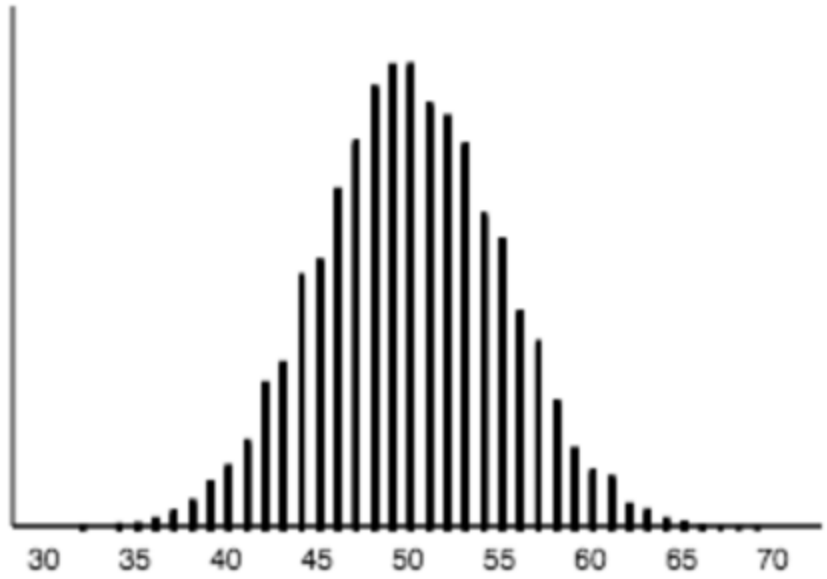
What is a sampling distribution?

Suppose we take many different random samples of 100 university students from a university that has an equal number of men and women.

The number of women will vary amongst the samples. For example, one sample could have 45 women, another sample could have 48 women, another sample could have 52 women, etc.

Though it could be possible that we get a random sample that only has 2 women in it, it would be pretty unlikely. Instead, we would expect that most of the samples would have around 50 women in it with some variation around that value.

Figure 1 is the result of a simulation that took 10,000 samples of size 100 from a population that had an equal about of women and men. The horizontal axis is the number of women in each sample. The height of each bar is the number of samples that had that many women.



Number of women in each sample of size 100

Notice how the most common number of women is around 50 (i.e. the average), but there is variation from that 50. Most samples have between 40 and 60 women.

The variability among random samples of size *n* from the same population is called **sampling variability**.

A probability distribution that characterizes some aspect of sampling variability is termed a **sampling distribution**. A sampling distribution is constructed by taking all possible samples of a size *n* from a population. Then for each sample, a statistic is calculated (e.g. sample mean, sample proportion, sample standard deviation). The sampling distribution is then created by making a graph of all of these samples.

Actually constructing a sampling distribution is often very difficult. A medium sized university in Canada might have 12,000 students. All possible samples of size 100 from that population would result in 5.87×10^{249} unique samples! Think about that. One billion is 10^9. Google is named after a googol (10^{100}) because they wanted Google to be

associated with an immense amount of data. Yet a googol is smaller than all possible samples at 100 from the medium sized university. If we got a computer to find all possible samples, it would take it over a billion years to find them[footnote]! Therefore, actually constructing a true sampling distribution in most situations is incredibly hard, incredibly time consuming, and not really worth it. Thus when we talk about sampling distributions, we talk about a **theoretical sampling distribution**. That is, we theorize what this sampling distribution would look like if it was possible to examine all possible samples.
This is based off of calculations on how long it would take a network of supercomputer in 2011 to work through all possible combinations of a 256 bit encryption. By the way, there are less possibilities in a 256 bit encryption than there are all possible samples of size 100 from a population of 12,000.

Due to these limitations, we often look at an empirical sampling distribution, instead of a theoretical sampling distribution. An **empirical sampling distribution** is created by taking many samples from a population and finding a statistic for each sample, but not doing this for all possible samples. The plot shown in is an example of an empirical sampling distribution as it only contains 10,000 samples and not all possible samples. The statistic in is the number of women, but we could have also looked at the proportion of women.

In summary, a sampling distribution is a distribution of a statistic. This differs from other distributions, like the population distribution, which are distributions for individual data values.

Why do we care about sampling distributions?

Suppose we take a random sample of 100 students from a medium sized university and we find that 75 of them are women. Does this call into question the assumption that 50% of the students are women? This is hard to figure out unless we know how likely it is that we could have found this random sample, assuming that there are an equal number of men and women.

The sampling distribution helps us find this probability. From the empirical sampling distribution in Figure 1 we can find the probability of getting a random sample of 75 women, assuming that there are an equal number of men and women is 0.0000%. That is, it is really unlikely to get a random sample of 75 women out of 100 if there are an equal number of men and women in the population. Based on this, we can be fairly confident that this university probably doesn't have an equal number of men and women. Instead, it is more likely that there are women than men at this university.

The process described above is called **inferential statistics**. Inferential statistics is used to make a conclusion about the population (all students at the university) from a sample (100 students). In general, to do any form of inferential statistics, we need to use a sampling distribution to either determine how likely or unlikely a statistic is (in hypothesis testing) or to estimate a parameter from a statistic (confidence intervals).

Thus sampling distributions are the backbone of inferential statistics.

Note: What was described above about the proportion of women at a university should sound familiar. In Chapter 4, we used the binomial distribution to determine how not unlikely or unlikely events were. The binomial distribution was helping us understand the sampling distribution of proportions.

Constructing empirical sampling distributions - MRU - C Lemieux
Introduction to how to construct a sampling distribution.

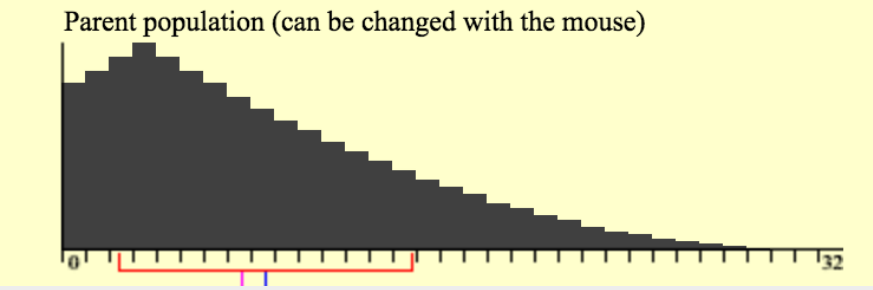
How to construct an emprical sampling distribution

If we have access to the population, we can construct an empirical distribution from it. This can be done by using computer software to pull random samples from a population. An example of one such tool is from the Rossman Chance website, which has an applet that allows you to create an empirical sampling distribution from a finite population: http://www.rossmanchance.com/applets/OneSample53.html

When constructing an empirical sampling distribution, it is important to keep the law of large numbers in mind. That is, the more samples you take, the closer the empirical sampling distribution will be to the theoretical sampling distribution. In general, empirical sampling distributions should be constructed from at least 10,000 samples.

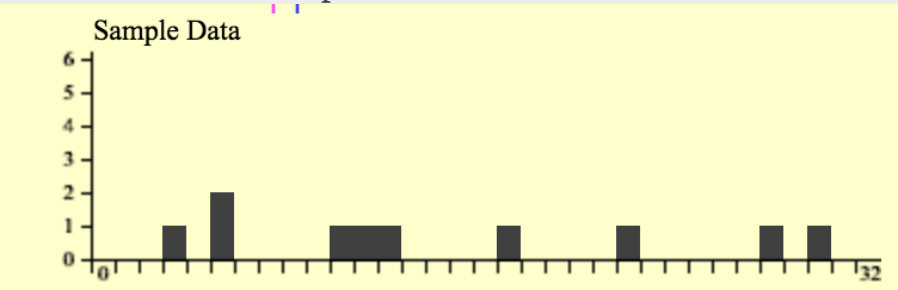
To get an idea of how an empirical sampling distribution is constructed, go to http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Example:
The images/figures in this example were generated from David Lane's sampling distribution applet that is part of the OnlineStatBook project [footnote].
Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
Figure 1 shows the histogram of the population we are going to generate an empirical sampling distribution from. We call this population the **parent population** as it is the population we are creating the sampling distribution from. Notice that the parent population is skewed left.



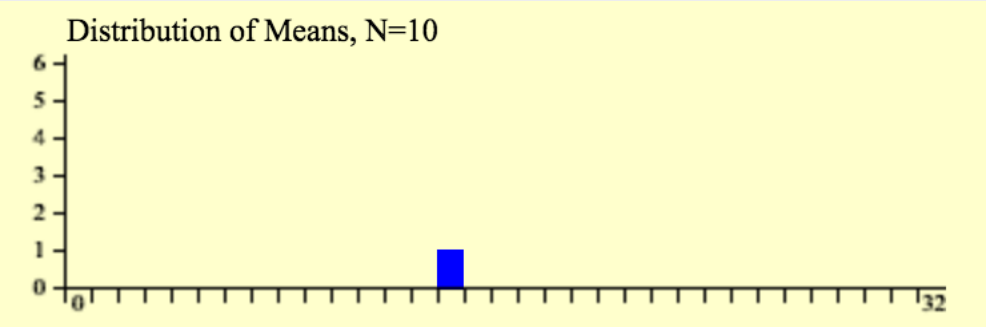
Parent population

We are going to take multiple samples of size 10 from the parent population and look at the statistic of the sample mean for each sample. Here is the first sample:



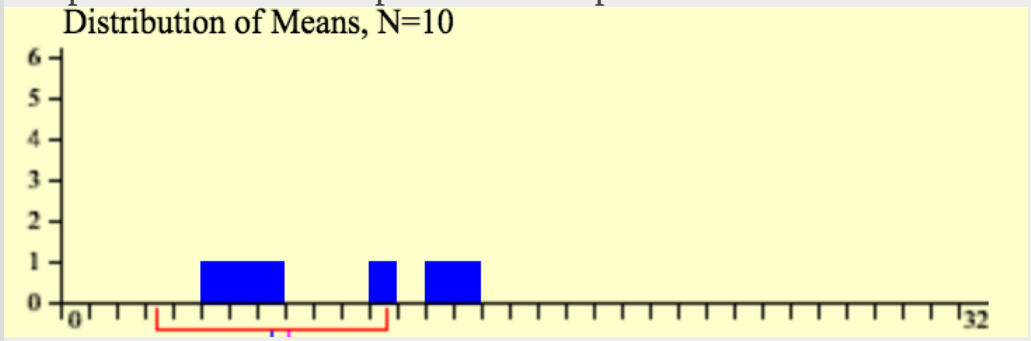
Sample of size 10 from the parent population

This is the sample mean of the sample:



Sample mean for one sample of size 10

Now one sample mean is not enough to tell us what the sampling distribution looks like. So let's take a few more samples. Let's take 5 more samples of size 10 and plot their sample means:

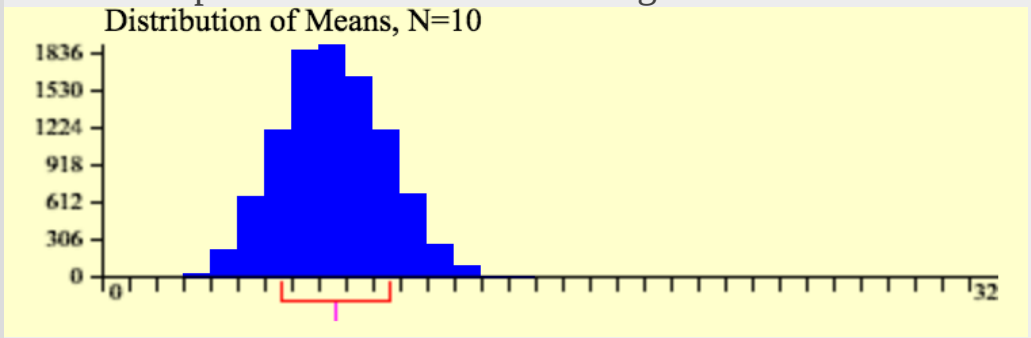


Six sample means from parent population

This is still a pretty small sample.

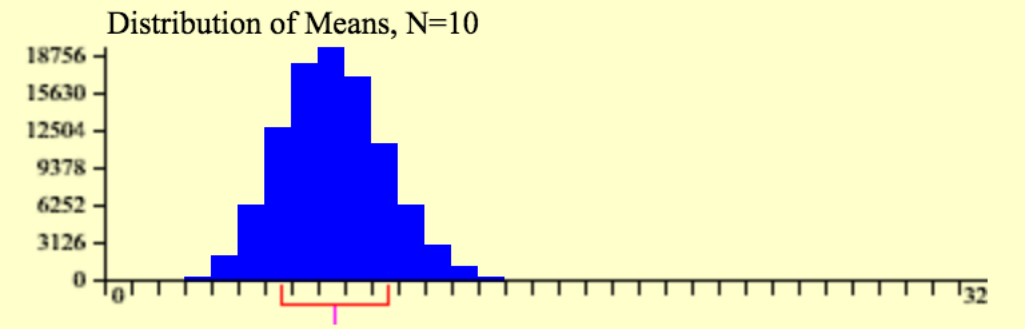
Note: There are two sample sizes here. One is the size of the sample we are taking from the parent population (10). The other is the number of samples we've taken (6). The first is the sample size for the sample. The second is the sample size for the empirical sampling distribution.

Now let's take 10,000 samples of size 10 from the population and plot each of their sample means. This is what we get:



10,000 sample means from parent population

Finally, let take 100,000 samples of size 10 from the population and plot each of their sample means. This is what we get:



100,000 sample means from parent population

Notice how there is no real difference between the distributions (shape, centre and variation) in Figure 5 and Figure 6. This means that are empirical distribution is now giving us a good sense of what the theoretical sampling distribution would look like. When this happens, this is called **convergence**. That is, the empirical sampling distribution is converging on the theoretical sampling distribution. As the sample size of the empirical sampling distribution increases this is expected to happen due to law of large numbers.

Bootstrapping

Suppose we don't have access to the population. This can happen if the population is infinite (e.g. in a manufacturing process) or where the population is large (e.g. population of Canada) or where most researchers wouldn't have access to the population (e.g. list of students at a university). Can we still construct an empirical sampling distribution?

The answer is yes! To do this, we use a process called **bootstrapping**. Essentially bootstrapping follows the same procedure as outlined in Example 1, but instead of using a parent population, we use a parent

sample. That is, we take a good sample from the population and use that to construct the sampling distribution.

Again the law of large numbers applies. If the random sample from the population is large enough, then the sample will most likely be a good estimate of the population. Then the empirical sampling distribution generated by the sample will most likely be a good estimate of the theoretical sampling distribution of the population.

Note: Bootstrapping only works if the sample being used has been collected properly and that the sampling technique ensures that the sample is random, the sample is representative of the population, and the sample size is large enough. There are no set rules on how big the sample needs to be, but for bootstrapping the bigger the better.

Central Limit Theorem - MRU - C Lemieux
Introduction to the key properties of the central limit theorem for the mean and proportions.

Another way to determine what the sampling distribution looks like is by using theory. The main theory that helps us understand the characteristics of the sampling distribution is called the **central limit theorem**.

The central limit theorem is an incredibly useful and powerful theorem. The theorem tells us about the distribution of many different sampling distributions. But be careful! The central limit theorem cannot be applied always and only applies to sampling distributions.

The central limit theorem for the sampling distribution for sample means

The sampling distribution for the sample means comes from a parent population that is comprised of quantitative data. Random samples of size *n* are taken from the parent population and the sample mean is calculated for each sample. What will the distribution of the sample means look like? That is, what is the shape of the distribution of sample means, where are the sample means centred, and what is the sampling variability?

Note: The following refers to the theoretical sampling distribution for the sample means. Further, when sample size is mentioned, it is referring to the size of the sample taken from the population. That is, it is not referring to how many different random samples have been taken.

Where are the sample means centred?

As the sample means are estimating the population mean, it makes sense that the sample means are centred around the population mean.

In the previous section, we saw the right skewed parent population in Figure 1. The population mean of that parent population is 8.08. Notice that the empirical sampling distributions shown in Figures 5 and 6 are both centred around 8.08.

In general, the mean of the theoretical sampling distribution for the sample means equals the population mean.
Equation:

$$\mu_x = \mu_x$$

Note: The variable for the sample means is x. That is why the subscript for the mean of the sample means (μ_x) has changed.

What is the sampling variability? (or what is the variation in the sampling distribution)

Based on the law of large numbers, the sampling variability of the sample means will decrease as the sample size increases. As the sample size increases, the sample means will become better and better estimates of the population mean and, therefore, there will be less variability between them. That is, there will be more variability between the sample means for samples of size 2, then there will be for samples of size 30.

Just like we can measure variability for individual data values, we can also measure variability for sample means. We will use the standard deviation to measure the sampling variability. The standard deviation of the sampling distribution for sample means is called the **standard error of the sample means**. It is found with the following formula [footnote] :
This formula assumes the population is infinite or very large. If this is not the case, then the formula is
Equation:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \quad \sqrt{\frac{N-n}{N-1}}$$

As the population size (*N*) increases, $\frac{N-n}{N-1}$ approaches 1 and no longer affects the standard error.

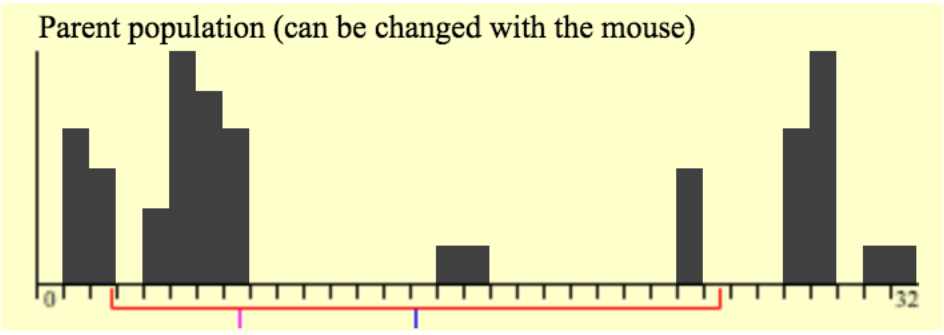
Equation:

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

What is the shape of the distribution?

This is actually a really interesting question.

Suppose the parent population looks like this [footnote]:
The images/figures that follow were generated from David Lane's sampling distribution applet that is part of the OnlineStatBook project Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.



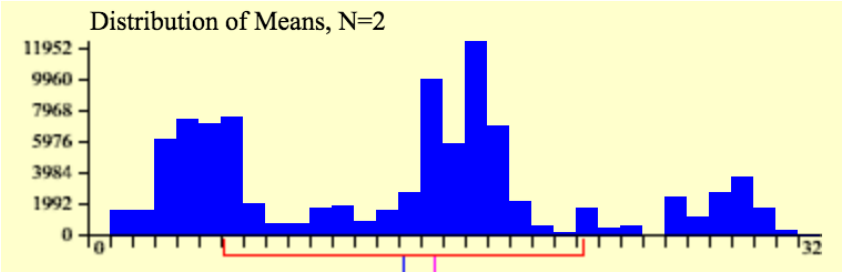
Parent population

What will the sampling distribution for sample means look like?

Here's the answer:

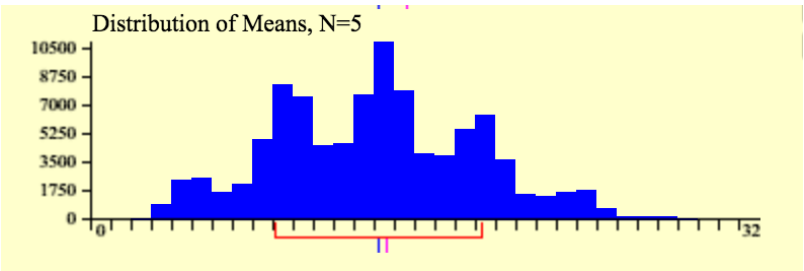
- If the parent population is normal, then the sampling distribution for sample means will be normal. Always.
- As the sample size of the samples being taken from the parent population increases, the more normal the sampling distribution for sample means will become.

Since the population in Figure 1 is not normally distributed, then we would expect the sampling distribution will not be normal for smaller sample sizes, but will be normal for larger sample size.

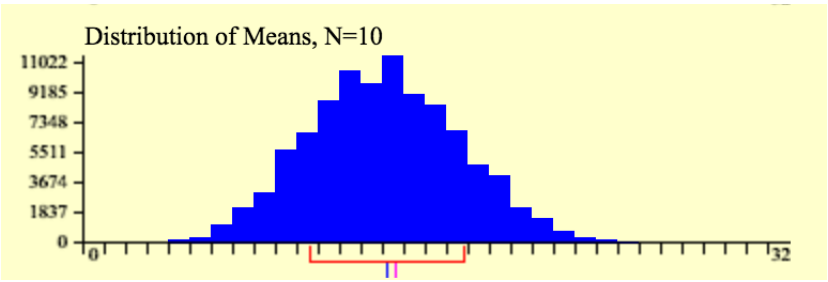


Sampling distribution for Figure 1 for samples of size 2

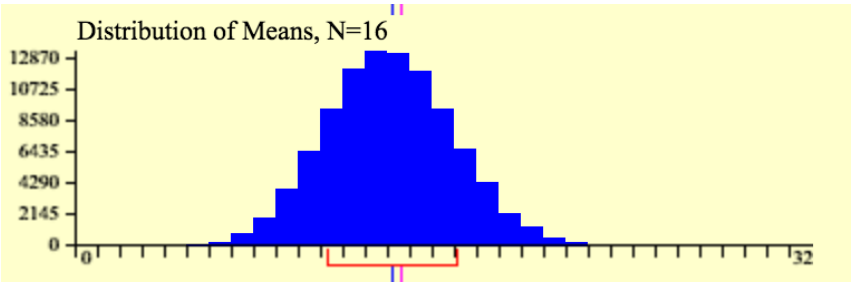
Note: For each of these empirical sampling distributions, 100,000 samples were taken of size **n**. Therefore, we can be very confident that the empirical sampling distributions are good representations of the theoretical sampling distributions.



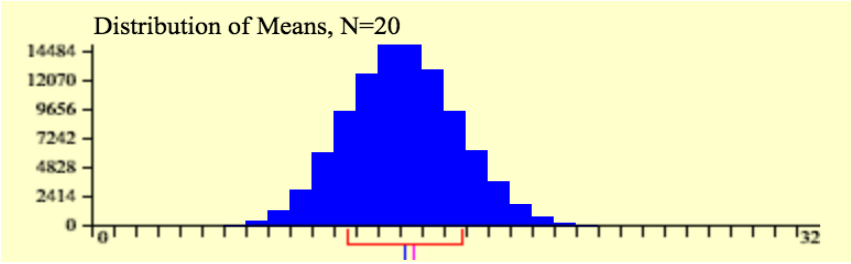
Sampling distribution for Figure 1 for samples of size 5



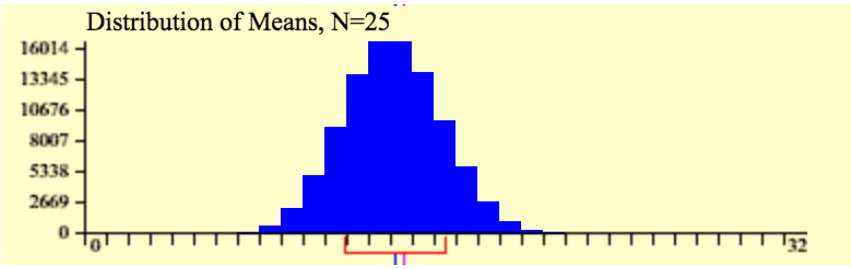
Sampling distribution for Figure 1 for samples of size 10



Sampling distribution for Figure 1 for samples of size 16



Sampling distribution for Figure 1 for samples of size 20



Sampling distribution for Figure 1 for samples of size 25

Figure 1 (the parent population) is not even close to being normal, but notice that as the sample size increases, the sampling distribution for sample means gets closer and closer to being normally distributed!

In general, the closer the population is to being normally distributed, the "faster" the sampling distribution gets closer to normal. Here "faster" means for a smaller sample size.

Note: The central limit theorem states that regardless of the shape of the population, if the sample size is greater than 30, the sampling distribution will be approximately normal.

Measure	Population	Sample	Sampling distribution for the sample mean
Mean	μ_x	x	$\mu_x = \mu_x$
Standard deviation	σ_x	s_x	$\sigma_x = \frac{\sigma}{\sqrt{n}}$ (standard error)

Summary of measures

The central limit theorem for the sampling distribution for sample proportions

The sampling distribution for the sample proportions comes from a parent population that satisfies the criteria of the binomial distribution. Random samples of size *n* are taken from the parent population and the sample proportion is calculated for each sample. What will the distribution of the sample means look like? That is, what is the shape of the distribution of sample proportions, where are the sample proportions centred, and what is the sampling variability?

The sampling distribution for sample proportions has similar characteristics as the sampling distribution for the sample means.

Where are the sample proportions centred?

They are centred around the population proportion.

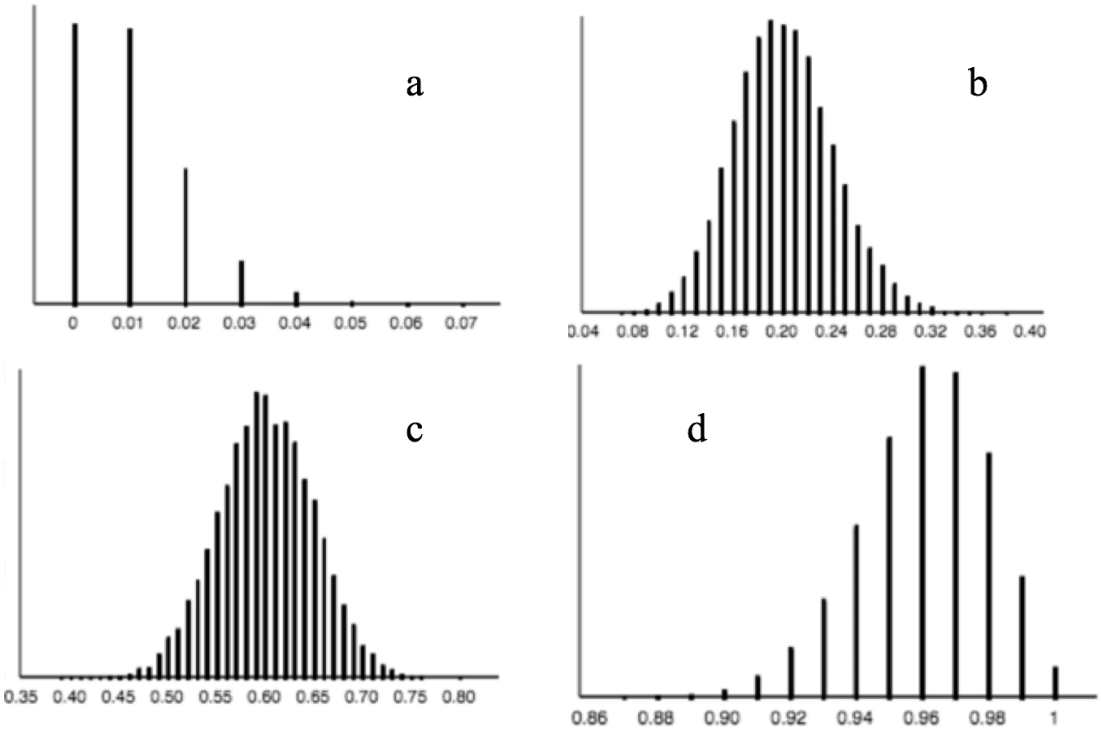
What is the sampling variability?

It decreases as the sample size increases.

What is the shape?

The shape of sampling distributions of the sample proportions also becomes normal. Unlike for sample means though, the normality is not based on sample size, but is based on the number of successes ($n\pi$) and failures ($n\left(1-\pi\right)$).

To illustrate, here are the empirical sampling distributions for proportions for various population proportions. The sample size is 100 in each case and the number of samples taken is 10,000.



Empirical sampling distributions for sample proportions

In Figure 8 a, n =100 and π = 0.01. Therefore, the number of successes is 1 and the number of failures is 99. The sampling distribution is skewed to the right.

In Figure 8 b, n =100 and π = 0.20. Therefore, the number of successes is 20 and the number of failures is 80. The sampling distribution is approximately normal.

In Figure 8 c, n =100 and π = 0.60. Therefore, the number of successes is 60 and the number of failures is 40. The sampling distribution is approximately normal.

In Figure 8 d, n =100 and π = 0.96. Therefore, the number of successes is 96 and the number of failures is 4. The sampling distribution is skewed to

the left.

Note: In general, the shape of the sampling distribution for sample proportions is approximately normal if the number of successes and the number failures are both at least 5.

If the sampling distribution for sample proportions is normal, we can find probabilities for the distribution using two methods. The first method is using the binomial distribution. The second method is the normal distribution. This might seem a bit strange as the binomial distribution is for discrete random variables and the normal distribution is for continuous random variables. In reality, we use the normal distribution to approximate probabilities for the sampling distribution for sample proportions. This is called the **normal approximation to the binomial distribution**. To get the exact probability, one would need to use the binomial distribution. But this can be cumbersome when the sample sizes are very large (e.g. 1000). Therefore, using the normal distribution can be beneficial, especially because it gives very accurate approximations. In example 6.4 below we will investigate this further.

Further when we begin to do inferential statistics, we won't know the population proportion (otherwise inferential statistics wouldn't be necessary). Since we won't know π it will hard to use the binomial distribution. Therefore, we use the normal approximation to the binomial distribution instead.

If we use a normal approximation to the binomial distribution, we need to know the mean and standard deviation of the sampling distribution.

The mean of the sampling distribution for sample proportions is the population proportion.
Equation:

$$\mu_{\hat{p}} = \pi$$

The standard deviation of the sampling distribution for sample proportions (or the **standard error of sample proportions**) is found using the following formula:

Equation:

$$\sigma_{\widehat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

A series of examples - Calculating probabilities for sampling distributions - MRU - C Lemieux
Four examples that show how to calculate and interpret probabilities found for sampling distributions

If a sampling distribution is normally distributed, then we can find probabilities for the sampling distribution using the normal distribution just like we did in Chapter 5.

The z-score for the sampling distribution for sample means would be:
Equation:

$$Z = \frac{X - \mu_x}{\sigma_x} = \frac{X - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

The z-score for the sampling distribution for sample proportions would be:
Equation:

$$Z = \frac{\widehat{p} - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}} = \frac{\widehat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Exercise:

Problem:

The Old Baldy Tire Company is introducing a new steel belted radial tire. Old Baldy's engineering department has estimated that the mean life of this tire will be 50,000 km and that the standard deviation of these tires will be 10,000 km. Suppose a large number of random samples of 100 tires is taken. The shape of the population distribution is unknown.

1. Can we assume the distribution of the mean life of these tires will be normal? Explain.
2. Regardless of your result in a), assume that we are dealing with a normal distribution. Find the probability that the mean life of a

random sample of 100 tires is less than 49,000km.

3. A competitor of Old Baldy's takes a random sample of 100 tires and finds their mean life to be 49,000 km. Based off of this data, they claim that the engineering department of Old Baldy's has exaggerated the mean life of their new tires. Do you support the competitor's claim? Explain.

Solution:

a. Yes. As the sample size is greater than 30 (it is 100), we can assume that the sampling distribution of the sample mean lifespan of the tires is normally distributed regardless of the shape of the sampling distribution due to the central limit theorem.
b. x = mean lifespan of 100 Old Baldy tires, $\mu_x = \mu_x$ = 50,000, $\sigma_x = \frac{\sigma}{\sqrt{n}}$ = 10,000/10 = 1000. Since we know that the data is normally distributed, we can use a computer program to calculate the probability $\mathrm{P}(X < 49{,}000)$. From the computer program, we get $\mathrm{P}(X < 49{,}000)$= 15.87%
c. No. The probability that a random sample of 100 Old Baldy tires has a mean lifespan of 49,000km is 15.87% (assuming Old Baldy's claim). This means that this event is likely to occur (as it is greater than 10%), under the assumption that the tires last on average 50,000 km, and does not provide sufficient evidence against Old Baldy's claim.

Exercise:

Problem:

The maintenance manager at a popular mountain resort is trying to determine if the aging gondola is in need of some renovation—or perhaps outright replacement. Right now, the maximum load of the gondola is 900 kilograms or 12 persons. The manager knows that the average weight of North Americans has been on the rise for several years and wishes to test what the probabilities might be of this gondola exceeding the maximum capacity.

Since the operators don't currently look at gender—just numbers—the manager is concerned about what might happen if the worst-case scenario were to occur: 12 large adult males were allowed on the gondola at the same time.

To investigate this further the manager did some research into the current average weight of adult males and discovered that it is about 80 kilograms. He also knows that adult weight tends to be normally distributed by gender, with a standard deviation for males of about 12 kilograms.

a. Given this information, he first wants to know what the individual weight allowance is (i.e. the per person average) that the gondola can withstand.
b. He also wants to know how likely is it that the individual weight of any randomly selected male will exceed the individual weight allowance calculated above.
c. Finally, he wants to know how likely it would be that the average weight of a sample of 12 adult males would exceed the average individual weight allowance.
d. Based on your answers, do you think the manager should renovate the gondola? Is there any further information that the manager would need?

Solution:

a. 900kg/12 people = 75kg/person
b. Since this is about an individual, we will use μ_x and σ_x. As stated in the question, we know that the population is normally distributed. From this, a computer program calculated that $\mathrm{P}(\mathrm{X} \geq 75)$ = 66.15%, where *X* is the weight of an individual person on the gondola.
c. Now, we are being asked about the mean for 12 people. Therefore, this question is about finding a probability for the sampling distribution for sample means. Therefore, we will use x = mean weight of 12 people, $\mu_x = \mu_x$ = 80, $\sigma_x = \frac{\sigma}{\sqrt{n}}$ = 12/3.46

= 3.46. Since the population distribution is normal, we know that the sampling distribution will also be normal (regardless of the sample size). Therefore, we can use a computer program to calculate the probability $\mathrm{P}(X > 75)$. We get 92.55%.

d. The probability found in c) is the probability that the average mass of 12 adult males will exceed the maximum individual weight for the gondola. The next question is "how likely is it that there will be 12 adult males on the gondola?" The manager should do further research to determine this before making a decision. While waiting for the results, the manager should implement a policy where any large groups of males are broken up and are required to take the lift in separate groups. I.e. break up a group of 12 males into two groups of 6 males.

Exercise:

Problem:

The city of Montreal has an extensive bike lane system. In fact, it is one of the largest in North America. But many cyclists find that even with all of the bike lanes, it is still hard to get around the city on a bike. In particular, there are many lanes that run east/west, but few that run north/south. Thus, they are encouraging the city council to focus less on adding lots of kilometers to the system, but instead making sure that the current system properly connects all parts of the city.

The city council will only go forward with this idea if at least 66% of the residents support focusing on connecting the system rather than expanding the system.

Suppose that 62% of residents do support connecting the system rather than expanding it. What is the probability that a random sample of 1000 residents will have a sample proportion of at least 66%?

a. Find the above probability using the binomial distribution.
b. Find the above probability by using the sampling distribution for sample proportions.
c. Compare the two answers. Do they give similar answers?

d. Based on your answers, do you think that it is possible that the city of Montreal will choose to focus on connecting the bike path system?

Solution:

a. Since we are using the binomial distribution, we are being asked to find the probability that at least 660 of the 1000 people in the poll will want to focus on connecting the system. The 660 comes from 66% of 1000. In other words, we are asked to find the $\mathrm{P}(\mathrm{X} \geq 660)$, with *n* = 1000 and π = 60%. Using a computer program, this yields a probability of 0.48%. This is found, by highlighting all of the values above 660 and including 660.

b. Since we are using the sampling distribution for sample proportions, we are asked to find the probability that the sample proportion will be at least 66%. In other words, we are asked to find the $P(\widehat{p} \geq 0.66)$. We can assume the sampling distribution for sample proportions is normal as the number of successes ($n\pi$ =1000 $\times$ 0.62=620) and the number of failures ($n\left(1-\pi\right)$ = 1000 $\times$ 0.38=380) are both at least five. Therefore, we will use the normal distribution to find the probability with $\mu_{\widehat{p}} = \pi$ = 0.62 and $\sigma_{\widehat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.62(1-0.62)}{1000}}$ = 0.01475. Therefore, using a computer program we find $P(\widehat{p} \geq 0.66)$ = 0.33%

c. The two probabilities are quite close. They are only 0.15% apart. Therefore, the two methods give us similar results.

d. It is unlikely that if the proportion of residents that want to focus on connecting the bike system is 62% that a poll of 1000 people would result in a sample proportion of 66%. Therefore, it is unlikely that the city of Montreal will chose to focus on connecting the system.

Exercise:

Problem:

Video games are gaining more and more popularity. Children often try to convince their parents to buy games even when they are not appropriate. For example, they may want to play a very violent game that is not appropriate for their age group. To help parents out, video games have rating categories to suggest age appropriateness. But how aware are parents of these categories?

To investigate this, you conduct a survey of Canadian families that have young children who play video games. You show parents three video game covers that have the category rating clearly marked on it. You then ask the parents whether the games would be appropriate for children and why. If the parent correctly identifies which games are appropriate for their children and refers to the ratings in making their choice, you categorize the parents as well informed.

Suppose that we want to use your results to justify the claim that less than 30 percent of parents are well informed about video game ratings. In your random sample of 1000 parents, you actually found that 27 percent of the parents that you polled were well informed about video game ratings.

a. Assuming that the proportion of parents that are well informed about video game ratings is 30%, what is the probability that you would observe a sample proportion of less than 27%. Use the normal approximation of the binomial distribution to find your answer.
b. Based on your results, do you believe that this is enough evidence to suggest that less than 30% of parents are well informed about video game ratings? Explain your answer.

Solution:

a. Since we are using the sampling distribution for sample proportions, we are asked to find the probability that the sample

proportion will be at most 27%. In other words, we are asked to find the $P(\widehat{p} \leq 0.27)$. We can assume the sampling distribution for sample proportions is normal as the number of successes (and the number of failures (($n\pi$ =1000 $\times$ 0.30=300) and the number of failures ($n\left(1-\pi\right)$ = 1000 $\times$ 0.7=700) are both at least five. Therefore, we will use the normal distribution to find the probability with $\mu_{\widehat{p}} = \pi$ = 0.30 and $\sigma_{\widehat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$ = $\sqrt{\frac{0.3(1-0.3)}{1000}}$ = 0.01145. Therefore, using a computer program we find $P(\widehat{p} \leq 0.27)$ = 0.44%.

b. Since the probability that we would observe a sample proportion of 27% (assuming a population proportion of 30%) is 0.44%, it is very unlikely we would have observed this evidence if the assumption is true. Therefore, it is more likely that the population proportion is less than 30%. Thus there is enough evidence to suggest that less than 30% of parents are well informed about video game ratings.

Practice questions

The following practice questions are from Lyryx Learning, Business Statistics I -- MGMT 2262 -- Mt Royal University -- Version 2016 Revision A. OpenStax CNX. Sep 8, 2016 http://cnx.org/contents/f3aefa9e-58d2-41ea-969f-04dc2cb04c82@5.5

Use the following information to answer the next ten exercises: A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely so the distribution of weights is uniform. A sample of 100 weights is taken. The standard deviation is 0.58 pounds.

Exercise:

Problem:

a. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deivation?

b. What is the distribution for the mean weight of 100 25-pound lifting weights?
c. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

Solution:

a. Uniform with a mean of 25 and a standard deviation of 0.58 pounds. Remember when a distribution is uniform all of the values are equally likely. Therefore the mean will be halfway between the lowest value (24) and the highest value (26).
b. Normal with a mean of 25 and a standard deviation of 0.0577
c. 0.0416

Exercise:

Problem:

Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

Solution:

0.0003

Exercise:

Problem:

Find the 90^{th} percentile for the mean weight for the 100 weights.

Solution:

25.07

Exercise:

Problem:

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. What is the probability that the 49 balls traveled an average of less than 240 feet?
b. Find the 80^{th} percentile of the distribution of the average of 49 fly balls.

Solution:

a. 0.0808
b. 256.01 feet

Exercise:

Problem:

According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

a. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
b. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

Solution:

a. Yes. I would be surprised, because the probability is almost 0.

b. No. I would not be totally surprised because the probability is 0.2312

Exercise:

Problem:

Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let $\bar{X}$ the average of the 49 races.

a. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
b. Find the 80^{th} percentile for the average of these 49 marathons.
c. Find the median of the average running times.

Solution:

a. 0.6247
b. 146.68
c. 145 minutes

Exercise:

Problem:

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, the mean of $\bar{X}$ is approximately equal to the mean of X.
b. When the sample size is large, $\bar{X}$ is approximately normally distributed.
c. When the sample size is large, the standard deviation of $\bar{X}$ is approximately the same as the standard deviation of X.

Solution:

a. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
b. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
c. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of X as the sample size increases.